Consistent estimates for categorical data based on a mix of administrative data sources and surveys

Boeschoten, Laura

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Consistent estimates for categorical data based on a mix of administrative data sources and surveys

Laura Boeschoten

Consistent estimates for categorical data based on a mix of administrative data sources and surveys

PhD thesis. Department of Methodology and Statistics, School of Social and Behavioral Sciences, Tilburg University, the Netherlands

ISBN:	978-94-6323-803-8
Author:	Laura Boeschoten
Cover Design:	Karin Aalberts
Printing:	Gildeprint

Consistent estimates for categorical data based on a mix of administrative data sources and surveys

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. K. Sijtsma, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op

vrijdag 25 oktober 2019
om 13.30 uur

door

Laura Boeschoten

geboren op 29 juni 1988
te Hilversum

Promotores:	prof. dr. A.G. de Waal
	prof. dr. J.K. Vermunt
Copromotor:	dr. D.L. Oberski
Promotiecommissie:	prof. dr. R.H.H. Groenwold
	prof. dr. L.A. van der Ark
	prof. dr. W. Smits
	dr. D. Vidotto

Manuscripts based on the studies presented in this thesis

Chapter 2 **Laura Boeschoten**, Daniel L. Oberski, Ton de Waal. Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent Class modeling (MILC). *Journal of Official Statistics*, 2017; 33(4): 921-962

Chapter 3 **Laura Boeschoten**, Ton de Waal, Jeroen K. Vermunt. Estimating the number of serious road injuries per vehicle type in the Netherlands using Multiple Imputation of Latent Classes (2019) *Journal of the Royal Statistical Society Series A: (Statistics in Society)*

Chapter 4 **Laura Boeschoten**, Danila Filipponi, Roberta Varriale. Combining Multiple Imputation and Latent Markov modeling to obtain consistent estimates of *true* employment status. *Journal of Survey Statistics and Methodology* Resubmitted after revisions

Chapter 5 **Laura Boeschoten**, Daniel L. Oberski, Jeroen K. Vermunt, Ton de Waal. Updating Latent Class Imputations with external auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal* 25(5): 750-761

Chapter 6 **Laura Boeschoten**, Marcel A. Croon, Daniel L. Oberski. A note on applying the BCH method under linear equality and inequality contraints (2018) *Journal of Classification*, 1-10

Chapter 7 **Laura Boeschoten**, Sander S. Scholtus, Jacco Daalmans, Ton de Waal, Jeroen K. Vermunt. Using Multiple Imputation of Latent Classes (MILC) to construct population census tables with data from multiple sources. *Unpublished*

Contents

PART I

Introducing the MILC method

Chapter 1

Introduction

Statistics Netherlands (or in Dutch 'Centraal Bureau voor de Statistiek', CBS) is the National Statistical Institute (NSI) of the Netherlands. The goal of CBS is to provide reliable statistical information on all sorts of aspects of the Dutch society (CBS, 2018a). Traditionally, CBS produced the statistics by using information collected by means of surveys. Generally, the surveys were conducted using a sample of the population. In addition, complete enumerations were held on a more incidental basis for the census (Nordholt et al., 2004). However, since the last quarter of the 20^{th} century, response rates to surveys were declining (De Heer & De Leeuw, 2002). To adapt to this issue, CBS changed the focus from producing statistics based on surveys to producing statistics using already available administrative data sources as much as possible. When producing these register-based official statistics, information from multiple registries is often needed simultaneously. For example, to produce a statistic describing the employment rate per gender requires unit-linked information from the BRP registry ('BasisRegistratie Personen', National Office for Identity Data, Ministry of the Interior and Kingdom Relations, 2016) and the POLIS registry (CBS, 2018b).

The statistics produced by CBS need to meet a number of requirements. First, for the majority of the statistics produced, international agreements have been made on how they should look exactly. This is for example the case with the census (The Economic and Social Council, 2005). The international agreements on the census regard the variables listed, the categories of which these variables consist and which cross-tables are produced between these variables (European Commission, 2008, 2009 and 2010). International agreements like these are important, as it can become very difficult to compare countries on specific topics otherwise. Second, to prevent any form of confusion, all statistics produced should be numerically consistent. This means that for example the total number of males listed when disseminating a statistic on employment rate per gender on a specific time-point is exactly equal to the total number of males in a statistic on level of education per gender on that same time-point.

To facilitate the production of official statistics meeting these requirements, a system of interlinked and standardized registers has been developed, the SSD (System of Social Statistical Datasets, Bakker et al., 2014), which is used for all social and spatial statistics produced by CBS. As the international agreements sometimes require production of statistics that are not found in the population registries, this information is acquired from sample surveys, which are therefore also included in the SSD.

However, when the information required to produce statistics of interest is collected from

the registries and surveys, it cannot be assumed that all observed measurements contain true scores. In a registry, errors are induced for example when the person filling out a registry simply makes a typo or checks a wrong box. Such a situation is not remarkable, as this person has other user purposes than a researcher or CBS has. Furthermore, categories used at the registry might not exactly match the definitions used by researchers or CBS. At last, with registries, it often happens that the same values keep being assigned until someone changes it. For example, when an employee resigns from his/her job, the employer should change this persons' status in the employment registry. If the employer forgets to do this, the status of this person will remain 'being employed', and this is also the status that is used when statistics are produced using this registry (Bakker, 2009; Groen, 2012).

Also for surveys we can think of many reasons why errors can be induced. In one example, you can think of a situation where a very specific number is asked, such as a persons' gross monthly income. In such cases, people often do not directly know the exact number, and are not willing to make the effort to find it out exactly. Therefore, they might try to retrieve the exact number from their memory, which might not be exactly the same as their true gross monthly income (Moore et al., 2000). In a second example, a respondent might be inclined to give social desirable answers when questioned about a sensitive topic (e.g. voting behavior) in a face-to-face setting with an interviewer (Loosveldt, 2008, p.215). In a third example, a respondent is filling out a web survey about expenditures, and figures out that certain initial responses prevent the appearance of more follow-up questions (e.g. answering 'yes' to the question 'Did you buy any clothing in the previous month?' might result in a number of follow up questions about what types of clothing you bought; Kreuter et al., 2011). To save time, the person might adjust his/her answers in such a way that fewer follow up questions will appear and thereby the given answers will deviate from the true scores.

When producing official statistics, the presence of misclassification, or measurement error, in data is problematic as it causes bias in both descriptives statistics of single variables, and in relationships between multiple variables. To illustrate the problem of misclassification, consider a situation where we have a multinomially distributed variable of interest X, which has K categories with corresponding probabilities $\mathrm{P}_1, ..., \mathrm{P}_K$. X can for example measure *type of employment contract*. Now imagine that X is not observed itself directly, but by variable Y in an administrative source. Y would then also measure *type of employment contract*. Furthermore, Y might not be a perfect measurement of X, i.e. Y contains misclassification. This is for example caused when a person has a permanent employment contract (the score of this person on variable X measuring *true type of employment contract*), but is listed as having

a temporary contract in the administrative source (the score of this person on variable Y measuring observed *type of employment contract*). We can then calculate $\mathrm{P}(Y)$ as

$$\mathrm{P}(Y) = \sum_{k=1}^{K} \mathrm{P}(X)\mathrm{P}(Y \mid X).$$

From here, it follows that generally

$$\mathrm{P}(Y) \neq \mathrm{P}(X),$$

which means that if official statistics are produced using observed variable Y, the statistics can contain bias.

In addition, when Y contains misclassification and is therefore not a perfect measurement of X, bias can also be present when statistics are produced about relationships with other variables. For example, if researchers are interested in the relationship between employment type and gender (measured by variable Q), then $\mathrm{P}(Y, Q)$ will also contain bias and is therefore not a perfect representation of $\mathrm{P}(X, Q)$.

As CBS has the aim to produce statistics of high quality, the measurements obtained from the surveys and registries cannot be used directly, as the misclassification should be corrected for. Sometimes, this can be done in a relatively straightforward manner. For example, when companies provide their monthly turnover to the Dutch tax office, they are requested to report it in multiples of $1,000$ Euro. If a respondent does not read the instructions carefully, he or she might provide it in Euro directly (De Waal et al., 2009, p.29). Such types of errors stand out, can be detected and it is clear how they can be corrected. Other types of errors can be detected because they result in an impossible combination with scores on other variables. An example is when a person scores 'married' on the variable *marital status* and 'under five years old' on the variable *age*. This specific combination of scores is logically impossible, it is even prohibited by law. However, it is not sure which of the two variables contains the error. Methods to handle errors that lead to impossible combinations, but for which the error mechanism is unknown, are often based on the Fellegi-Holt paradigm (Fellegi & Holt, 1976). Conceptually, with the Fellegi-Holt paradigm, first the edit rules defining the non-allowed score combinations should be specified. Second, each record in the dataset is made consistent with the edit rules by adjusting the smallest possible number of observed values.

Let us now consider a variable measuring *employment status* (denoted as Y_1) observed

in a registry. This variable might contain some misclassification that is not detected by the previously mentioned methods. Because of this misclassification, it differs from the *true* variable, measuring *true employment status* (X). Let us now consider a situation where a variable measuring *employment status* is also observed by means of a survey (and denote this by Y_2). This variable can also contain misclassification. However, the causes for misclassification in surveys are generally considered to be different from the causes for misclassification in registries. Therefore we can assume that the probability of misclassification in Y_1 is unrelated to the probability of misclassification in Y_2 (i.e. there is local independence). Furthermore, in such situations, CBS is often able to link the survey variables and register variables on person level using unique identifiers.

Such a person-linked combined dataset offers many possibilities when it comes to estimating and correcting for misclassification. If the combined dataset contains multiple locally independent variables measuring the same construct (such as *employment status*), statistical models such as latent variable models can be used to estimate a variable (X) using imperfect measurements (Y_1 and Y_2 in this example). In addition, under certain assumptions information can be obtained about the quality of the observed variables (Y_1 and Y_2), which is not possible when only one measurement is present. If X measures *true employment status*, it can be considered a categorical latent variable and a latent class (LC) model can be used to estimate X (Biemer, 2011, p.13). See also Figure 1.1 for an illustration.

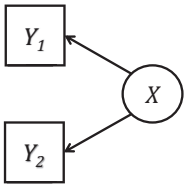

Figure 1.1: Basic set-up of the latent class model used to estimate *true employment status* as latent variable X by indicators Y_1 and Y_2. Note that this model in itself is not identifiable and is only shown for clarification.

LC analysis is traditionally considered as a method to identify a categorical latent variable using categorical observed variables (McCutcheon, 1987, p.7). Here, the latent variable generally corresponds to a target phenomenon which can only be measured indirectly. An example where LC analysis is applied in this traditional way is to estimate a latent variable

'eating disorder types' with the specific disorders as LC's and using indicators such as being underweight, excessive exercising and vomiting (Keel et al., 2004). A key assumption made here is that the probability of obtaining a specific marginal response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y})$ is a weighted average of the C class specific probabilities $\mathrm{P}(\mathbf{Y} = \mathbf{y}|X = x)$:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} \mathrm{P}(X = x)\mathrm{P}(\mathbf{Y} = \mathbf{y}|X = x). \tag{1.1}$$

Here, $\mathrm{P}(X = x)$ denotes the proportion of units belonging to category x in the underlying true measure. In the traditional LC example measuring eating disorder types, x is a specific eating disorder type such as *bulimia*. If LC analysis is used to correct for measurement error as we propose, for example to estimate the *true* variable measuring *employment status*, x is a score on the variable *true employment status* such as *unemployed*.

A second key assumption of LC modeling is that the observed indicators are independent of each other given an individual score on the latent variable. This implies that it is assumed that the relationships between the observed variables result from the fact that each of them is affected by the latent variable (McCutcheon, 1987). This is called the assumption of local independence and can be expressed as follows:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}|X = x) = \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l|X = x), \tag{1.2}$$

where L stands for the number of indicator variables. When considering this assumption in our combined dataset setting, it means that it is assumed that the probability of misclassification in the first indicator variable is independent of the probability of misclassification in the second indicator variable. As previously discussed, this assumption is likely to be met in cases where one indicator originates from a registry and another originates from a survey, as the causes for misclassification in surveys and registries are different. The same holds for situations where variables are used from different independently collected registries or different independently collected surveys. However, careful attention must always be paid to whether the different sources are collected independently and whether a person might purposely provide error prone answers in multiple sources. In these situations, the local independence assumption does not hold.

Combining equation 1.1 and equation 1.2 yields the following model for response pattern

$\mathrm{P}(\mathbf{Y} = \mathbf{y})$:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} \mathrm{P}(X = x) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x).$$

The model parameters ($\mathrm{P}(X = x)$ and $\mathrm{P}(Y_l = y_l \mid X = x)$) can then be estimated by Maximum Likelihood (ML) and this is implemented in statistical software such as Latent GOLD (Vermunt & Magidson, 2013a) or the poLCA package in R (Linzer et al., 2011). When used in such a setting, $\mathrm{P}(Y_l = y_l \mid X = x)$ can be used to obtain information on how response patterns of the indicators are related to scores of the 'true variable'. Furthermore, $\mathrm{P}(Y_l = y_l \mid X = x)$ provides information on the error rates of the single categories within the indicator variables and $\mathrm{P}(X = x)$ provides information on the distribution of the 'true variable' (Biemer, 2011, p.22).

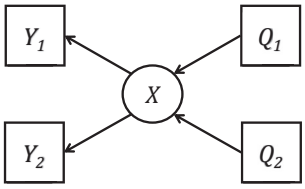

Figure 1.2: Basic set-up of the latent class model used to estimate *true employment status* as latent variable X by indicators Y_1 and Y_2. Two covariates, Q_1 and Q_2 are also included in the LC model.

If a researcher or CBS is interested in the relationship between X and other variables measured by either the survey or registry, these variables can be included in the LC model as covariates (denoted by $\mathbf{Q}$):

$$\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q}) = \sum_{x=1}^{C} \mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x). \tag{1.3}$$

The model now contains an additional set of parameters, $\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})$, providing information on how scores on the 'true variable' may vary between different groups of individuals (Biemer, 2011, p.22). In addition, if there are impossible combinations of scores between X and $\mathbf{Q}$, this can be made explicit when specifying the model. That is, by restricting the $\mathrm{P}(X \mid \mathbf{Q})$ concerned to be equal to zero.

To summarize, the LC model provides information on the distribution of X, information on

how response patterns on indicators are related to scores of X and information on how X may vary between different groups of individuals. However, the model does not provide us an observed variable containing the true scores of X, in a way that is flexible for further analysis or for production of graphs or tables. For example, generating a cross-table containing observed frequencies for (X, Q_1) is not possible using the LC model output alone.

To be able generate such output, in this thesis I propose to use multiple imputation (see e.g. Rubin (1987)) to assign estimates of the true scores of X to a variable, that we denote as W. The imputations are generated by sampling from the latent classes using the corresponding posterior membership probabilities, which are obtained by applying Bayes' rule to the latent class model:

$$\mathrm{P}(X = x \mid \mathbf{Y} = \mathbf{y}, \mathbf{Q} = \mathbf{q}) = \frac{\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x)}{\sum_{x=1}^{C} \mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x)}. \tag{1.4}$$

The posterior membership probabilities provide us the probability that a unit truly belongs to a certain latent class given its combination of scores on the indicators and covariates (i.e. its profile) and, for each profile, the posterior membership probabilities sum up to one. By sampling from the posterior membership probabilities m times and thereby generating m imputations of W, the magnitude of the differences between these m imputations reflect the uncertainty about what value to impute (Van Buuren, 2012, p.16) and thereby uncertainty about the *true* values of X.

If we are then interested in for example the observed frequencies for (X, Q_1), the next step would then be to estimate this cross-table from each of the m imputations of W. As the m versions of W are different, the estimated parameters will also be different. When the m parameter estimates are then pooled (using the well-known 'Rubin's rules' (Rubin, 1987, p.76)), estimated variance is a combination of the conventional sampling variance (within-imputation variance) and an extra variance component caused by the measurement error in the indicator variables (between-imputation variance) (Van Buuren, 2012, p.17).

The aim of this thesis is to introduce a new method that combines latent class analysis and multiple imputation to correct for misclassification in combined datasets. Furthermore, the thesis aims to evaluate the performance of this method under a range of different situations. In addition, it is investigated if and how well the method is able to handle specific practical issues occurring while compiling statistics from combined datasets, as having a general and flexible method that can be tailored to specific practical situations would be essential for its

relevance for CBS.

Thesis outline

The remainder of this thesis is organized as follows. In **Chapter 2**, the solution proposed in the previous section (Multiple Imputation of Latent Classes, MILC) is explained in more detail. It is shown step-by-step how the method can be applied and what assumptions are made when applying the method. Furthermore, the performance of the method is investigated by means of a simulation study using a relatively small set-up. Here, the MILC method is also applied to a combined survey-registry dataset from CBS.

In **Part II** of this thesis, it is investigated how the MILC method can be adapted to more realistic settings. More specifically, in **Chapter 3**, the MILC method is extended in such a way that it is able to estimate the number of serious road injuries per *vehicle type* in the Netherlands on a yearly basis using two incomplete registries. Multiple covariates are also included in the model, to be able to stratify the number of serious road injuries per *vehicle type* into relevant subgroups. Here, the model is extended using a 'quasi-latent variable' to be able to create multiple imputations for the covariate *region of accident*. This enables us to estimate the number of serious road injuries per *vehicle type* per *region*. In **Chapter 4**, the MILC method is extended by using latent Markov modeling, so that it can handle longitudinal data and correspondingly create multiple imputations for multiple time-points. As recently many researchers have investigated the use of latent Markov modeling to estimate *true employment status* using a combined dataset consisting of the Labor Force Survey (LFS) and population registries, a simulation study used to investigate the performance of this extension is based on this specific example. Furthermore, the extended MILC method is applied to a combined LFS-registry dataset from Italy.

Then in **Part III** of this thesis, the MILC method is extended on a different level. In **Chapter 5** the situation is discussed where latent variable imputations are used in further analyses with covariates. These relationships can only be correctly estimated if the covariates are included in the LC model. Otherwise, point estimates will be biased. To handle situations where covariates are not included in the LC model, the MILC method is extended in such a way that latent variable imputations can be updated to be conditional on the external covariates. To enable this, the MILC method is combined with two alternative approaches to 'three-step' modeling, the so-called ML and BCH approach. Simulation studies are performed to investigate the performance of these extensions. In addition, in **Chapter 6**, the BCH

correction approach is placed in a framework of quadratic loss functions and linear equality and inequality constraints. In this way, it is prevented that the method results in inadmissible solutions with negative cell proportions, and marginals can be fixed to specific values so that edit restrictions can be incorporated.

In **Part IV**, it is evaluated whether and how the MILC method can indeed be used to produce consistent estimates for statistics produced using combined datasets. Therefore, in **Chapter 7**, it is evaluated whether the MILC method can produce consistent population statistics for the Dutch Population Census using multiple sources prone to misclassification. In **Chapter 8**, the opportunities and limitations of the MILC method in its current form are discussed. It is also discussed how the MILC method can be used in practice to produce official statistics. At last, suggestions for further improvement are made.

Chapter 2

Estimating classification error under edit restrictions in combined survey-register data using Multiple Imputation Latent Class modeling (MILC)

Abstract

Both registers and surveys can contain classification errors. These errors can be estimated by making use of a combined dataset. We propose a new method based on latent class (LC) modeling to estimate the number of classification errors across several sources while taking impossible combinations with scores on other variables into account. Furthermore, the LC model, by multiply imputing a new variable, enhances the quality of statistics based on the combined dataset. The performance of this method is investigated by a simulation study, which shows that whether the method can be applied depends on the entropy R^2 of the LC model and the type of analysis a researcher is planning to do. Finally, the method is applied to public data from Statistics Netherlands.

This chapter was published as: L. Boeschoten, D.L. Oberski, T. De Waal (2017) Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modeling (MILC) *Journal of Official Statistics* 33 (4), 921-962

2.1 Introduction

National Statistical Institutes (NSIs) often use large datasets to estimate population tables covering many different aspects of society. A way to create these rich datasets as efficiently and cost effectively as possible is by utilizing already available register data. This has several advantages. First, known information is not collected again by means of a survey, saving collection and processing costs as well as reducing the burden on the respondents. Second, registers often contain very specific information that could not have been collected by surveys (Zhang, 2012). Third, statistical figures can be published more quickly, as conducting surveys can be time consuming. However, when more information is required than already available, registers can be supplemented with survey data (De Waal, 2016). Caution is then advised as surveys likely contain classification errors. When a dataset is constructed by integrating information on micro level from both registers and surveys, we call this a combined or composite dataset. More information on how to construct such a combined dataset can be found in Zhang (2012) and B. F. M. Bakker (2010). Composite datasets are used by, among others, the Innovation Panel (Understanding Society, 2016), the Millennium Cohort Study (UCL Institute of Education, 2007), the Avon Longitudinal Study of Parents and Children (Ness, 2004), the System of Social Statistical Databases of Statistics Netherlands and the 2011 Dutch Census (Schulte Nordholt et al., 2014).

When using registers for research, we should be aware that they are collected for administration purposes so they may not align conceptually with the target and can contain process delivered classification errors. These may be due to mistakes made when entering the data, delays in adding data to the register (B. F. M. Bakker, 2009) or differences between the variables being measured in the register and the variable of interest (Groen, 2012). This means that both registers and surveys can contain classification errors, although originating from different types of sources. This assumption is in contrast to what many researcher assume, namely that either registers or surveys are error-free. To illustrate, Schrijvers et al. (1994) used registers to validate a postal survey on cancer prevalence, Turner et al. (1997) used Medicare claims data to validate a survey on health status and Van der Vaart & Glasner (2007) used optician database information to validate a telephone survey. In contrast, Jörgren et al. (2010) used a survey to validate the Swedish rectal cancer registry and Robertsson et al. (1999) used a postal survey to validate the Swedish knee arthroplasty register. Since neither surveys or registers are free of error, it is most realistic to approach them both as such. Therefore, we aim to develop a method which incorporates information from both to estimate the *true* value, without assuming that either one of them is error-free.

To distinguish between two types of classification errors we classify them as either visibly or invisibly present. Both types can be estimated by making use of new information that is provided by the combined dataset. Invisibly present errors in surveys or registers can be detected when responses on both are compared in the combined dataset. Differences between the responses indicate that there is an error in one (or more) of the sources, although it is at this point unclear which score(s) exactly contain(s) error. The name 'invisibly present errors' is given because these errors could not have been seen in a single dataset. They can be dealt with by estimating a new value using a latent variable model. To estimate these invisibly present errors using a latent variable model, multiple indicators from different sources within the combined data are used that measure the same attribute. This approach has previously been applied using structural equation models (B. F. M. Bakker, 2012; Scholtus & Bakker, 2013), latent class models (Biemer, 2011; Guarnera & Varriale, 2016; Oberski, 2015) and latent Markov models (Pavlopoulos & Vermunt, 2015). Latent variable models are typically used in another context, namely as a tool for analyzing multivariate response data (Vermunt & Magidson, 2004).

Covariates (variables within the combined dataset that measure something other than the attribute of interest) can help improve the latent variable model. Some errors can then be observed already when an impossible combination between a score on the attribute and a covariate is detected, which we define as a visibly present error. The name 'visibly present errors' is given here because (some of) these errors are visible in a single dataset. An example of a combination which is not allowed is the score "own" on the variable *home ownership* and the score "yes" on the variable *rent benefit*. Such an, in practice, impossible combination can be replaced by a combination that is deemed possible. Whether a combination of scores is possible and therefore allowed is commonly listed in a set of edit rules. An incorrect combination of values can be replaced by a combination that adheres to the edit rules. Different types of methods are used to find an optimal solution for different types of errors (De Waal et al., 2012). For errors caused by typing, signs or rounding, deductive methods have been developed by Scholtus (2009, 2012) . For random errors, optimization solutions have been developed such as the Fellegi-Holt method for categorical data, the branch-and-bound algorithm, the adjusted branch-and-bound algorithm, nearest-neighbour imputation (De Waal et al., 2011, pp. 115-156) and the minimum adjustment approach (Zhang & Pannekoek, 2015). Furthermore, imputation solutions, such as nonparametric Bayesian multiple imputation (Si & Reiter, 2013) and a series of imputation methods discussed by Tempelman (2007) can be used.

The solutions discussed two paragraphs above for invisibly present errors are not tailored to handle the invisibly and visibly present errors simultaneously, and they do not offer possibilities to take the errors into account in further statistical analyses; they only give an indication of the extent of the classification errors. In addition, uncertainty caused by both visibly and invisibly present errors is not taken into account when further statistical analyses are performed. An exception is the method developed by Kim et al. (2015), which simultaneously handles invisibly and visibly present errors using a mixture model in combination with edit rules for continuous data, and which has been extended by Manrique-Vallier & Reiter (2016) for categorical data. This method allows for an arbitrary number of invisible errors based on one file and one measurement, whereas we consider multiple linked files with multiple measurements of an attribute. Any method dealing with visibly or invisibly present classification errors should account for the uncertainty created by these errors. This can be done by making use of multiple imputations (Rubin, 1987), and has previously been used in combination with solutions for invisibly present errors (Vermunt et al., 2008) and visibly present errors (Si & Reiter, 2013; Manrique-Vallier & Reiter, 2013) .

We propose a new method that simultaneously handles the three issues discussed: it handles both visibly and invisibly present classification errors and it incorporates them both, as well as the uncertainty created by them, when performing further statistical analysis. By comparing responses on indicators measuring the same attribute in a combined dataset we allow the estimation of the number of invisibly present errors using a Latent Class (LC) model. Visibly present errors are handled by making use of relevant covariate information and imposing restrictions on the LC model. In the hypothetical cross table between the attribute of interest and the restriction covariate, the cells containing a combination that is in practice impossible are restricted to contain zero observations. These restrictions are imposed directly when the LC model is specified. To also take uncertainty created by the invisibly and visibly present errors into account when performing further statistical analyses, we make use of Multiple Imputation (MI). Because Multiple Imputation and Latent Class analysis are combined in this new method, the method will be further denoted as MILC.

In the following section, we describe the MILC method in more detail. In the third section, a simulation study is performed to assess the novel method. In the fourth section, we apply the MILC method on a combined dataset from Statistics Netherlands.

2.2 The MILC method

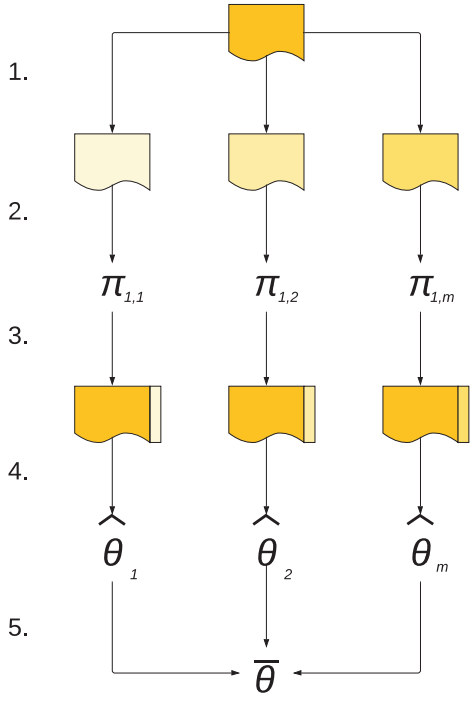

Figure 2.1: procedure of latent class multiple imputation for a multiply observed variable in a combined dataset.

The MILC method takes visibly and invisibly present errors into account by combining Multiple Imputation (MI) and Latent Class (LC) analysis. Figure 2.1 gives a graphical overview of this procedure. The method starts with the original combined dataset comprising L measures of the same attribute of interest. In the *first* step, m bootstrap samples are taken from the original dataset. In the *second* step, an LC model is estimated for every bootstrap sample. In the *third* step, m new empty variables are created in the original dataset. The m empty variables are imputed using the corresponding m LC models. In the *fourth* step, estimates of interest are obtained from the m variables and in the *fifth* step, the estimates are pooled using Rubin's rules for pooling (Rubin, 1987, p.76). These five steps are now discussed in more detail.

The MILC method starts by taking m bootstrap samples from the original combined dataset. These bootstrap samples are drawn because we want the imputations we create in a later step to take parameter uncertainty into account. Therefore, we do not use one LC model based on one dataset, but we use m LC models based on m bootstrap samples of the original dataset (Van der Palm et al., 2016).

In the next step, we make use of LC analysis to estimate both visibly and invisibly present classification errors in categorical variables. We first link several datasets by unit identifiers resulting in a combined dataset matched on a common core set of identifiers (discarding all records where no match is obtained), and group variables measuring the same attribute present on more than one of the original source datasets. For each of the variable groups we build a single latent variable (denoted by X) representing the underlining *true* measure, assuming discrepancies between different sourced measured.

For example, we have L dichotomous indicator variables ($Y_1, ..., Y_L$) measuring the same attribute *home ownership* (1="own", 2="rent") in multiple datasets linked on unit level. Differences between the responses of a unit are caused by what we described as invisibly present classification errors in one (or more) of the indicators. Since the indicators all have an equal number of categories (C), we fix the number of categories of the latent variable X to C.

The LC model we then build using the indicator variables is based on five assumptions. The first assumption pertains to the marginal response pattern $\mathbf{y}$, which is a vector of the responses to the given indicators. For example, we have three indicators measuring home ownership, the response pattern $\mathbf{y}$ can be "own", "own", "rent". We assume here that the probability of obtaining this specific marginal response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y})$ is a weighted average of the X class specific probabilities $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid X = x)$:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} \mathrm{P}(X = x)\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid X = x). \tag{2.1}$$

Here, $\mathrm{P}(X = x)$ denotes the proportion of units belonging to category x in the underlying *true* measure, where x might be "own", the proportion of the population owning their own house.

The second assumption is that the observed indicators are independent of each other given a unit's score on the underlying true measure. This means that when a mistake is made when filling in a specific question in a survey, this is unrelated to what is filled in for the same

question in another survey or register. This is called the assumption of local independence,

$$\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid X = x) = \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x). \tag{2.2}$$

Combining Equation 2.1 and Equation 2.2 yields the following model for response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y})$:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} \mathrm{P}(X = x) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x). \tag{2.3}$$

The model parameters ($\mathrm{P}(X = x)$ and $\mathrm{P}(Y_l = y_l \mid X = x)$) are estimated by Maximum Likelihood (ML). To find the ML estimates for the model parameters, Latent Gold uses both the Expectation-Maximization and the Newton-Raphson algorithm (Vermunt & Magidson, 2013a).

In Equation 2.3, only the indicators are used to estimate the likelihood of being in a specific true category. However, it is also possible to make use of covariate information to estimate the LC model. The third assumption we then make is that the classification errors are independent of the covariates. An example of a covariate which can help in identifying whether someone owns or rents a house is *marital status*, this covariate is denoted by Q and can be added to equation 2.3:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid Q = q) = \sum_{x=1}^{C} \mathrm{P}(X = x \mid Q = q) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x). \tag{2.4}$$

Covariate information can also be used to impose a restriction on the model, to make sure that the model does not create a combination of a category of the *true* variable and a score on a covariate that is in practice impossible. For example, when an LC model is estimated to measure the variable *home ownership* using three indicator variables and a covariate (denoted by Z) measuring *rent benefit*, the impossible combination of owning a house and receiving rent benefit should not be created.

Throughout the paper, we compare four approaches researchers might administer when performing analyses using combined datasets containing classification errors and edit restrictions. In the first approach, researchers completely ignore the combined data structure and directly use one variable (which measures a construct that is measured by other variables in the combined dataset as well) to obtain estimates of interest, e.g. a cross-table proportion or a logistic regression coefficient. In the second approach, researchers use an LC model to correct for classification errors, but are not aware of the edit restriction. The LC model used in

this approach is equal to Equation 2.4; we denote this as the *unconditional model*. In the third approach, researchers are aware of the edit restriction, but they assume that including the restriction covariate (Z) in the LC model is enough to account for this; they do not explicitly mention the restriction itself. We denote this as the *conditional model*:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid Q = q, Z = z) = \sum_{x=1}^{C} \mathrm{P}(X = x \mid Q = q, Z = z) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x). \tag{2.5}$$

Only in the fourth approach, the restriction is imposed directly in the LC model to fix the cell proportion of the impossible combination to zero; we denote this as the *restricted conditional model*. In the example where Z measures *rent benefit*, and the latent *true* variable measures *home ownership*, the imposed restriction is:

$$\mathrm{P}(X = \text{own} | Z = \text{rent benefit}) = 0. \tag{2.6}$$

By using such a restriction, we can take impossible combinations with other variables into account, while we estimate an LC model for the underlying *true* measure. The restriction is imposed by specifically denoting which cell in the cross-table between the covariate and the latent variable should contain zero observations and giving this cell a weight of exactly zero, resulting in constrained estimation (Vermunt & Magidson, 2013b).

By specifying a model as in Equation 2.4 or in Equation 2.5, we assume that the covariate measure is in fact error-free, which is the fourth assumption we make. A fifth assumption is that the edit rules applied are hard edit rules, in contrast to soft edit rules where there is a small probability that the edit is in fact possible. These five assumptions (assumption that $\mathrm{P}(\mathbf{Y} = \mathbf{y})$ is a weighted average of $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid X = x)$; assumption of local independence; assumption that classification errors are independent of covariates; assumption that the covariate is error-free; assumption of hard edits) are specific for the LC model we use.

However, in practice it is very likely that one of these assumptions is not met. For example, with the assumption of local independence, we assume that when a mistake is made in one indicator, this is unrelated to the answers on other indicators. This assumption is probably met when one indicator originates from a survey and another from a register. If two indicators both originate from surveys, it is much more likely that a respondent makes the same mistake in both surveys, this assumption would then not be met. We can also think of situations where the assumption that misclassification is independent of covariates is not met. For example with tax registrations by businesses, the number of delays and mistakes tends to be related to

company size, since appropriate administration is better institutionalized in larger companies. The assumption that a covariate is free of error is in practice almost never met, since all sources always contain some error. The last assumption made is that the edits applied are hard edits. In some cases soft edits might be more appropriate, for example when a combination of scores is highly unlikely but not impossible, such as the combination of being 10 years old and being graduated from high school.

Luckily these assumptions can be relaxed by specifying more complex LC models. However, whether you are able to relax these assumptions depends on your specific data structure. More specifically, it depends on whether your model is still identifiable. Unfortunately, model identifiability is not straightforward. For example, a model with three dichotomous indicators is identifiable, while a model with two dichotomous indicators is not. Adding a covariate to this model would make it identifiable. Adding a restriction to a model can also help to make an unidentifiable model identifiable. Since it is not possible to present general recommendations here, we refer to Biemer (2011) for more information about model identifiability. Examples of complex latent variable models which incorporate the different assumptions discussed in official statistics datasets are Pavlopoulos & Vermunt (2015) and Scholtus & Bakker (2013). Model identification can be checked in Latent Gold by assessing whether the Jacobian of the likelihood is full rank at a larger number of random parameter values (Forcina, 2008). All models in this paper were confirmed to be identifiable this way.

How missing values in the indicators and covariates are handled is also dependent on model specification. We specified the model as such that the indicators are part of the estimation procedure. Missing values are therefore handled by full information maximum likelihood (FIML) (Vermunt & Magidson, 2013b, pp.51-52). Covariates are treated as fixed and listwise deletion will be applied to missing values here.

By applying Bayes' rule to the LC models from Equation 2.4, Equation 2.5 or Equation 2.6, posterior membership probabilities can be obtained. These posterior membership probabilities represent the probability of being in an LC given a specific combination of scores on the indicators and covariates ($\mathrm{P}(X = x|Y = y, Q = q, Z = z)$). For example, the posterior membership probabilities for the *conditional model* are obtained by:

$$\mathrm{P}(X = x \mid Y = y, Q = q, Z = z) = \frac{\mathrm{P}(X = x \mid Q = q, Z = z) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x)}{\sum_{x=1}^{C} \mathrm{P}(X = x \mid Q = q, Z = z) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x)}.$$

These posterior membership probabilities can then be used to impute latent variable X. To

distinguish between the unobserved latent variable X, described by the LC model, and the variable after imputation, we denote this imputed variable by W. Different methods exist to obtain W. An example is modal assignment, where each respondent is assigned to the class for which its posterior membership probability is the largest. To correctly incorporate uncertainty caused by the classification errors, we use multiple imputation to estimate W. We first create m empty variables ($W_1, ..., W_m$) and we impute them by drawing one of the LCs by sampling from the posterior membership probabilities from the m LC models.

With the *restricted conditional model*, we wanted to make sure that cases were not assigned to categories on the latent *true* variable which would result in impossible combinations with scores on other variables, such as the combination "rent benefit" $\times$ "own". Therefore, the restriction set in Equation 2.6 is also used here.

After we created m variables by imputing them using the posterior membership probabilities obtained from each of the m LC models, the estimates of interest can be obtained. For example, we can be interested in a cross table between imputed *true* variable W and covariate Z, where our estimate of interest $\hat{\theta}$ can be the cell proportion P($W = 1, Z = 1$). The m estimates of $\hat{\theta}$ can now be pooled by making use of the rules defined by Rubin for pooling (Rubin, 1987, p.76). The pooled estimate is obtained by

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i.$$

The total variance is estimated as

$$\text{VAR}_{\text{total}} = \overline{\text{VAR}}_{\text{within}} + \text{VAR}_{\text{between}} + \frac{\text{VAR}_{\text{between}}}{m},$$

where $\overline{\text{VAR}}_{\text{within}}$ is the within imputation variance calculated by

$$\overline{\text{VAR}}_{\text{within}} = \frac{1}{m} \sum_{i=1}^{m} \text{VAR}_{\text{within}_i}.$$

$\text{VAR}_{\text{within}_i}$ is estimated as the variance of the proportion of $\hat{\theta}_i$,

$$\frac{\hat{\theta}_i \times (1 - \hat{\theta}_i)}{N},$$

where N is the number of units in the combined dataset, and $\text{VAR}_{\text{between}}$ is calculated by

$$\text{VAR}_{\text{between}} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})(\hat{\theta}_i - \hat{\theta})'.$$

Besides the uncertainty caused by missing or conflicting data represented by the spread of parameter estimate values, $\text{VAR}_{\text{between}}$ also contains parameter uncertainty, which was introduced by the bootstrap performed in the first step of the MILC method.

2.3 Simulation

2.3.1 Simulation approach

To empirically evaluate the performance of MILC, we conducted a simulation study using R (R Core Team, 2014). We start by creating a theoretical population using Latent Gold (Vermunt & Magidson, 2013a) containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) measuring the latent dichotomous variable (X); one dichotomous covariate (Z) which has an impossible combination with a score of the latent variable; and one other dichotomous covariate (Q). The theoretical population is generated using the restricted conditional model. When samples are drawn, it can happen that the LC model estimated from a sample assigns a non-zero probability to an impossible combination, so these errors are due to sampling. Furthermore, variations are made in the generated datasets according to scenarios described in the following paragraphs.

When evaluating an imputation method, the relation between the imputed latent variable and other variables should be preserved since these relations might be the subject of research later on. When investigating the performance of MILC, there are two relations we are particularly interested in. We are interested in the relation between the imputed latent variable W and the covariate Z, which has an impossible combination with a score on the latent variable. The four cell proportions of the 2×2 table are denoted by: $W1 \times Z_1$, $W_2 \times Z_1$, $W_1 \times Z_2$ and $W_2 \times Z_2$. The cell $W_1 \times Z_2$ represents the impossible combination, and should contain zero observations. We compare the cell proportions of a 2×2 table of the population latent variable X and Z with the cell proportions of a table of the imputed latent variable W and Z from the samples. Furthermore, we are interested in the relation between W and covariate Q. To investigate this relation, we compare the coefficient of a logistic regression of the latent population variable X on Q with the coefficient of the logistic regression of the imputed W on Q.

To investigate these relations, we look at three performance measures. First, we look at the bias of the estimates of interest. The bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the estimate over all replications. To confirm that the standard errors of the estimates are properly estimated, the ratio of the average standard error of the estimate over the standard deviation of the $1,000$ estimates was also examined.

We expect the performance of MILC to be influenced by the measurement quality of the indicators, the marginal distribution of covariates Z and Q, the sample size and the number of multiple imputations. The quality of the indicators is represented by classification probabilities. They represent the probability of a specific score on the indicator given the latent class. If the quality of the indicators is low, it will also be more difficult for MILC to assign cases to the correct latent classes.

From Geerdinck et al. (2014) we know that classification probabilities of 0.95 and higher can be considered realistic for population registers. Pavlopoulos & Vermunt (2015) detected a classification probability of 0.83 in the Dutch Labour Force Survey. We investigate a range of classification probabilities around the values found, from 0.70 to 0.99. The marginal distribution of Z, P(Z), is also expected to influence the performance of MILC. A higher value for P($Z = 2$) can give, for example, more information to the latent class model to assign scores to the correct latent class. Sample size may influence the standard errors and thereby the confidence intervals. The performance of MILC can also depend on the number of multiple imputations. Investigation of several multiple imputation methods have shown that 5 imputations are often sufficient (Rubin, 1987). However, with complex data, it can be the case that more imputations are needed. As a result, the simulation conditions can be summarized as follows:

- Classification probabilities: 0.70; 0.80; 0.90; 0.95; 0.99.
- P($Z = 2$): 0.01; 0.05; 0.10; 0.20.
- Sample size: $1,000$; $10,000$.
- Logit coefficients of X regressed on Q of $\log(0.45/(1-0.45) = -0.2007$, $\log(0.55/(1-0.55) = 0.2007$ and $\log(0.65/(1-0.65) = 0.6190$ corresponding to an estimated odds ratio of 0.81, 1.22 and 1.86. The intercept was fixed to zero.

- Number of imputations: 5; 10; 20; 40.

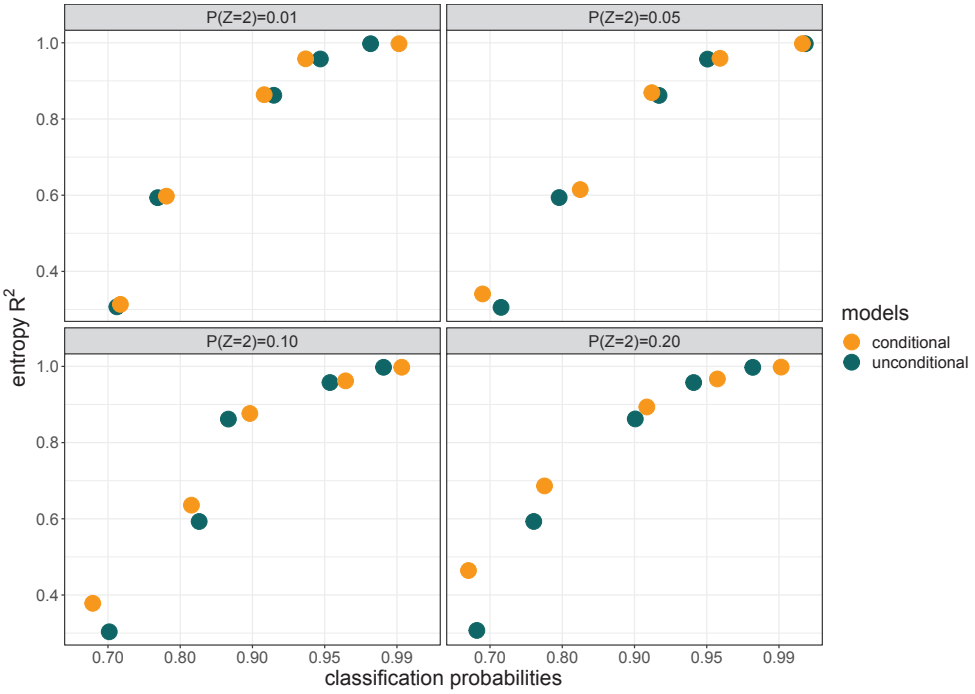

Figure 2.2: entropy R^2 of the unconditional and conditional model with different values for the classification probability and P($Z = 2$). The restricted conditional model has the same entropy R^2 as the conditional model because the models contain the same variables.

To illustrate the measurement quality corresponding to different conditions, Figure 2.2 shows the entropy R^2 of the models under different values for P($Z = 2$) and classification probabilities. The entropy indicates how well one can predict class membership based on the observed variables, and is measured by:

$$\text{EN}(\alpha) = -\sum_{j=1}^{N}\sum_{x=1}^{X}\alpha_{jx} log \alpha_{jx},$$

where α_{jx} is the probability that observation j is a member of class x, and N is the number of units in the combined dataset. Rescaled with values between 0 and 1, entropy R^2 is measured by

$$R^2 = 1 - \frac{\text{EN}(\alpha)}{N \log X},$$

where one means perfect prediction (Dias & Vermunt, 2008). The *conditional* and the *restricted conditional model* have the same entropy R^2 because these models contain the same variables.

All models with classification probabilities of 0.90 and above have a high entropy R^2 and are able to predict class membership well. When the classification probabilities are 0.70, the entropy R^2 is especially low. However, for the conditional and the restricted conditional model, the entropy R^2 under classification probability 0.70 increases as P($Z = 2$) increases. A larger P($Z = 2$) means that covariate Z contains more information for predicting class membership. Because covariate Z is not in the *unconditional model*, it makes sense that entropy R^2 remains stable for different values of P($Z = 2$) under this model. Furthermore, Figure 2.2 demonstrates that the performance of MILC is evaluated over an extreme range of entropy R^2 values and gives an indication of what we can expect from the MILC method under different simulation conditions.

2.3.2 Simulation results

In this section we discuss our simulation results in terms of bias, coverage of the 95% confidence interval, and the ratio of the average standard error of the estimate over the standard deviation of the estimates. We do this in three sections. In the first section we discuss the 2×2 table of the imputed latent variable W and restriction covariate Z. In the second section, we investigate the relation between the imputed latent variable W and covariate Q. In the third section we investigate the influence of m, the number of bootstrap samples and multiple imputations. In the simulation results discussed in the first two sections, we used $m = 5$. When investigating the different simulation conditions, we focus on the performance of the four approaches discussed, using one indicator (Y_1), the *unconditional model*, the *conditional model* and the *restricted conditional model*. Interesting findings are illustrated with graphs containing results from situations when Y_1 is used and W estimated using the restricted conditional model. For conditions that yielded approximately identical results, only one condition is shown in the figures. In Appendix A, tables with all results from the four approaches are given.

2.3.2.1 The relation of imputed latent variable W with restriction covariate Z

When we investigate the results in terms of bias (Figure 2.3), the restricted conditional model produces bias when the classification probabilities of the indicators are below 0.80. The bias of the cells where $P(Z = 1)$ for the restricted conditional model decreases when the classification probabilities increase or when P($Z = 2$) increases. This trend coincides with the trend we saw in Figure 2.2 for the entropy R^2, where a high entropy R^2 corresponds to a low bias. In contrast, when Y_1 is used, the bias of all cells is low when P($Z = 2$) is small, and increases

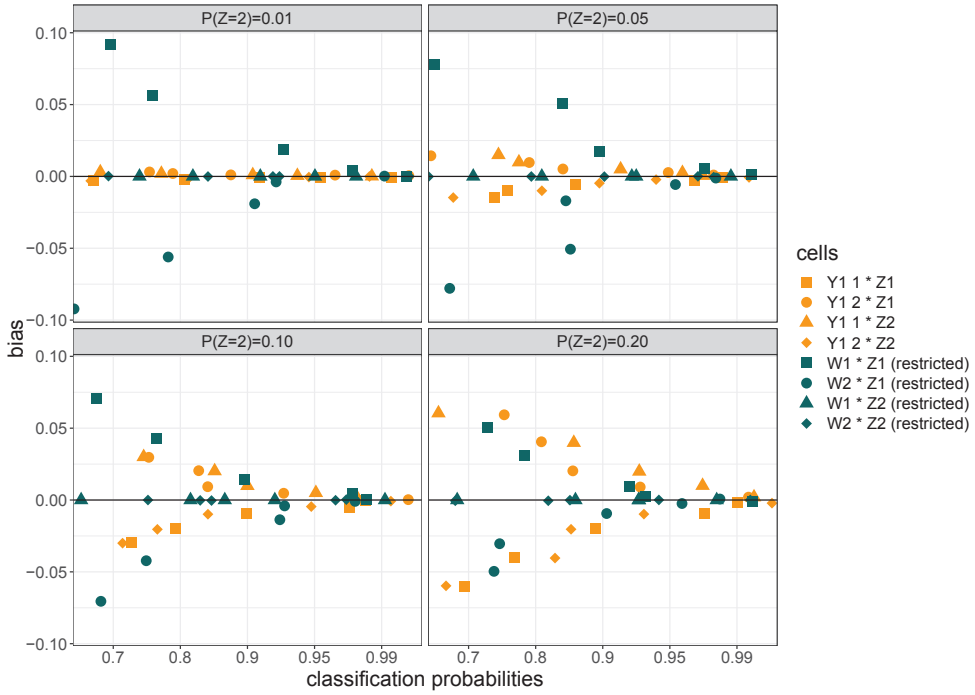

Figure 2.3: Bias of the four cell proportions of the 2×2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $\mathrm{P}(Z = 2)$. Sample size is $1,000$ and $m = 5$.

as $\mathrm{P}(Z = 2)$ increases. Furthermore, the *restricted conditional model* is the only model in which the cell representing the impossible combination $(W_1 \times Z_2)$ indeed contains zero observations. $(Y1_1 \times Z_2)$ is never exactly zero.

When investigating the results for coverage of the 95% confidence intervals around the cell proportions (Figure 2.4), we see that the results differ over the different sample sizes. This is caused by the fact that even though the bias is not influenced by the sample size, the standard errors and therefore the confidence intervals are. Confidence intervals of biased estimates are therefore less likely to contain the population value. Furthermore, if the classification probabilities are larger, individuals are more likely to end up in the correct latent class, which also results in less variance, resulting in smaller confidence intervals. Confidence intervals cannot be properly estimated for the impossible combination $Y1_1 \times Z_2$, since the proportions are very close to zero. This can be seen in Figure 2.4. Since $W_1 \times Z_2$ is not estimated with the *restricted conditional model*, confidence intervals cannot be estimated and coverage is therefore not shown.

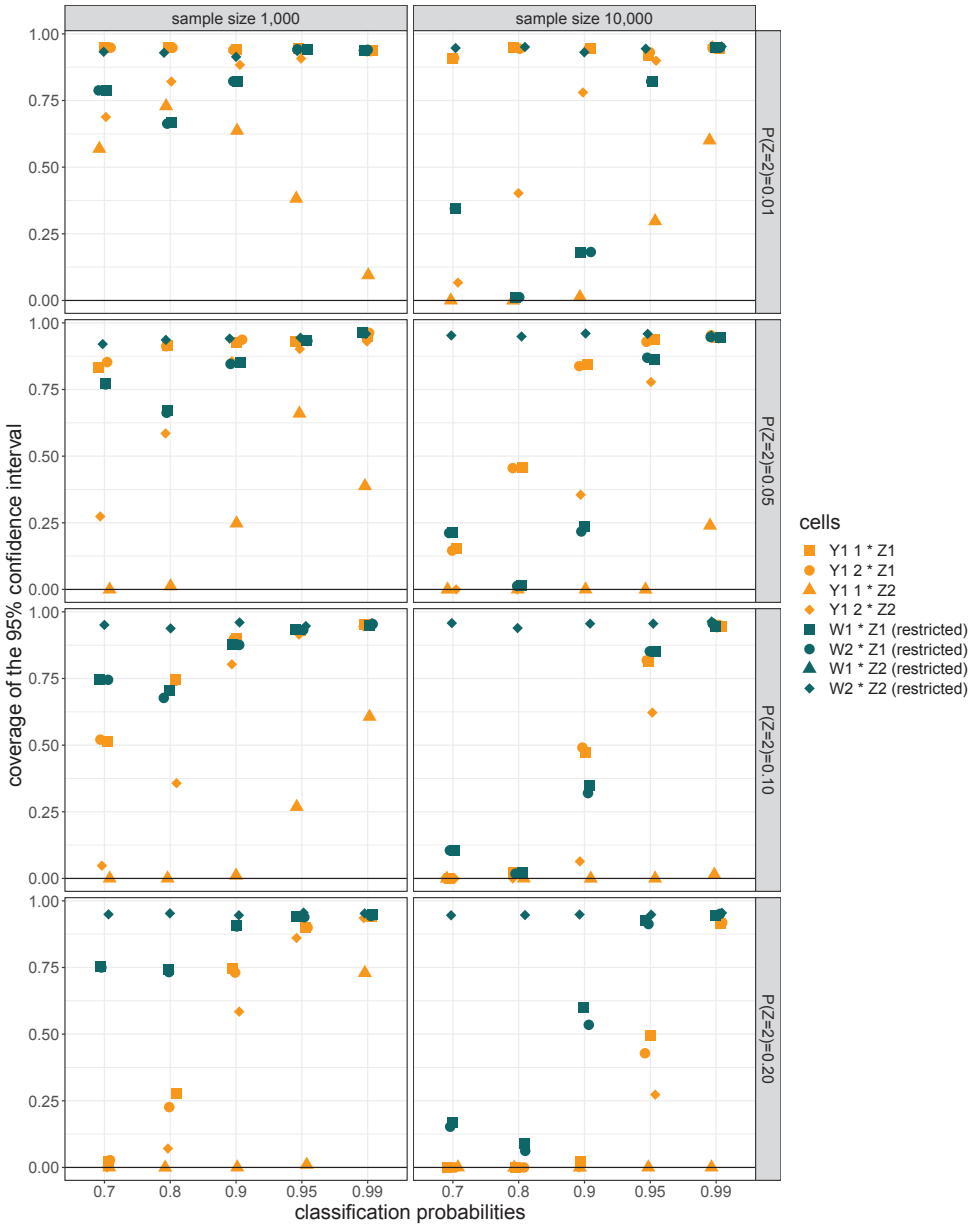

Figure 2.4: Coverage of the 95% confidence interval of the four cell proportions of the 2×2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $\mathrm{P}(Z = 2)$ and sample size. $m = 5$.

The ratio of the average standard error of the estimate over the standard deviation of the simulated estimates tells us whether the standard errors of the estimates are properly

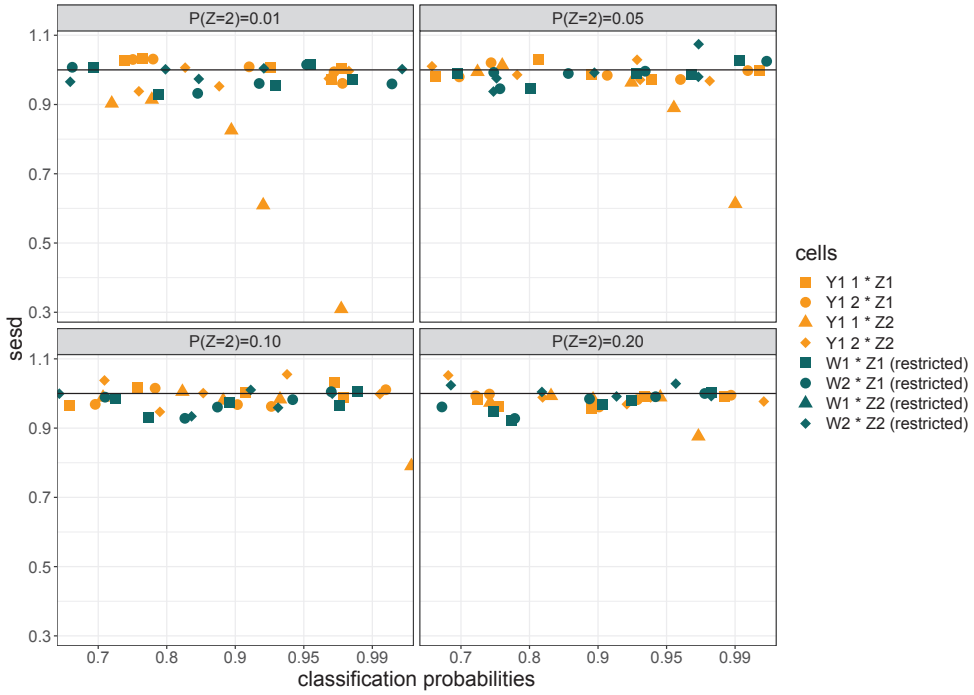

Figure 2.5: $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$. Sample size is $1,000$ and $m = 5$.

estimated. In general, the values for both the situation of one indicator and the restricted conditional model, found in figure 2.5 are both very close to one. Only the standard errors for $W_1 \times Z_2$ are too small when one indicator is used. With the *restricted conditional model*, these are not estimated.

Overall, the small 2×2 cross tables investigated here containing a restriction covariate can be estimated when the LC model of the combined dataset has an entropy R^2 of 0.90, or when the sample size is large, an entropy R^2 of 0.95.

2.3.2.2 Relationship between the imputed latent variable W and covariate Q

In the simulation results discussed in Section 2.3.2.1, the relation between the imputed latent variable W and covariate Z containing an impossible combination was investigated. Within the *restricted conditional model*, there was also another covariate, Q. We investigate the relation between W and Q with three different strengths of relations: intercepts are zero and logit

coefficients of W regressed on Q are -0.2007; 0.2007; 0.6190. Because the intercept is zero in all conditions, we focus on the coefficients of Q when investigating the simulation results.

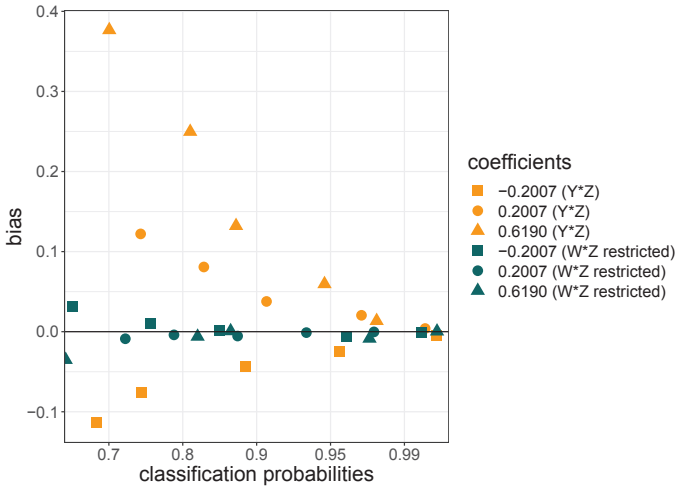

Figure 2.6: Bias of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient and the classification probabilities. $\mathrm{P}(Z = 2) = 0.01$, sample size is $1,000$ and $m = 5$.

In Figure 2.6 we see that for the restricted conditional model, the bias is very close to 0 in all conditions. When Y_1 is used, the bias is much larger and is related to the classification probabilities.

In Figure 2.7 we see the results in terms of coverage of the 95% confidence interval. The conclusions we can draw here are comparable to the conclusions we drew from the results in terms of bias. When W is used (estimated using the restricted conditional model), the coverage of the 95% confidence is approximately 95 in all discussed conditions. When only one indicator (Y_1) is used, we see undercoverage when the population value of the logistic regression coefficient is 0.6190. This undercoverage is related to the classification probabilities and increases when the sample size increases. Results in terms of the ratio of the average standard error of the estimate over the standard deviation of the simulated estimates are very close to the desired ratio of 1. This is the case for all investigated simulation conditions for both when Y_1 is used or when W is used. Results are reported in Appendix A.

Overall, for the investigated conditions, unbiased estimates can be obtained when the LC model of the combined dataset has an entropy R^2 of 0.60 or larger.

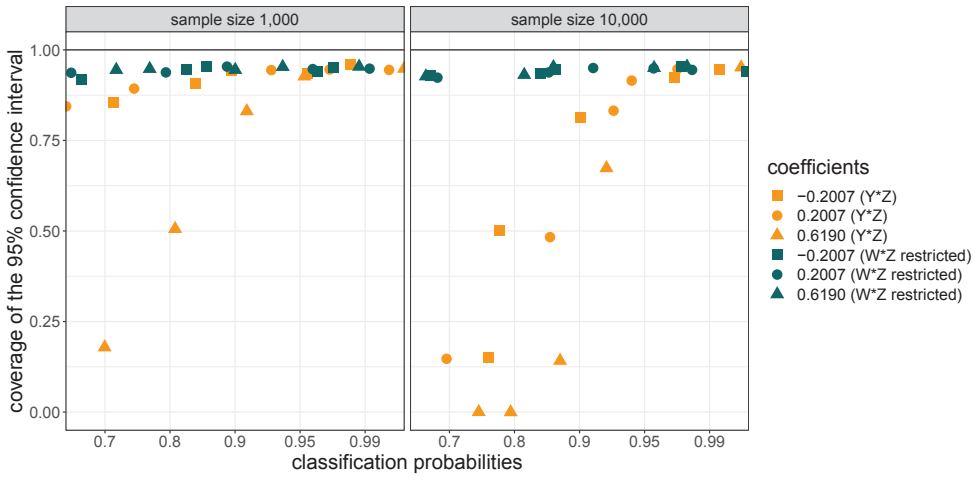

Figure 2.7: Coverage of the 95% confidence interval of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient, the classification probabilities and sample size. $\mathrm{P}(Z = 2) = 0.01$ and $m = 5$.

2.3.2.3 Number of imputations

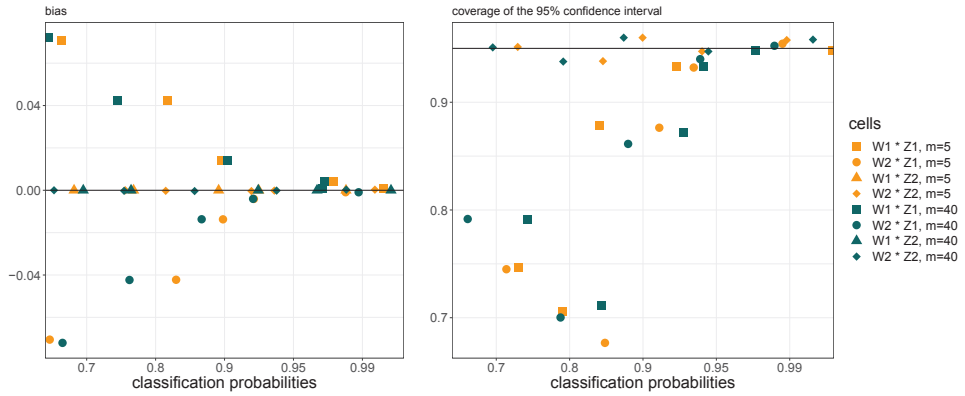

Figure 2.8: Bias and coverage of the 95% confidence interval of four cells in the 2×2 table of covariate $Z \times W$ (estimated using the *restricted conditional model*). Number of bootstrap samples $m = 5$ and 40. The sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.10$.

To investigate the effect of the number of bootstrap samples and imputations (m), we performed 5, 10, 20 and 40 bootstrap samples and imputations. The results of $m = 5$ and $m = 40$ can be found in figure 2.8, while more results can be found in Appendix A. Both in terms of bias and coverage the MILC method performs equally well over the different numbers of m. It is important to note that the fraction of missing information corresponds, in the worst case, to the amount of missing data (Rubin, 1987, p.114). In our case, it depends

on the entropy R^2, which is dependent on the classification and the covariates. Although the amount of missing values in W is 100%, the amount of missing information is much smaller when the entropy R^2 is larger than 0. This might explain why biased estimates with inappropriate coverage are obtained when the entropy R^2 is low, regardless of the size of m.

2.4 Application

2.4.1 Data

Home ownership is an interesting variable for social research. It has been related to a number of properties, such as inequality (Dewilde & Decker, 2016), employment insecurity (Lersch & Dewilde, 2015) and government redistribution (André & Dewilde, 2016). Therefore, we apply the MILC method on a combined dataset that brings together survey data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013 (Scherpenzeel, 2011), which is administered by CentERdata (Tilburg University, The Netherlands) and a population register from Statistics Netherlands from 2013. Because samples for LISS were drawn by Statistics Netherlands, we were very well able to link these sources. From this combined dataset, we use two variables indicating whether a person is either a home-owner or rents a house/other as indicators for the imputed *true* latent variable *home-owner/renter or other*. The combined dataset also contains a variable measuring whether someone receives rent benefit from the government. A person can only receive rent benefit if this person rents a house. In a cross-table between the imputed latent variable *home-owner/renter* and *rent benefit*, there should be zero persons in the cell "home-owner $\times$ receiving rent benefit". If people indeed receive rent benefit and own a house, this could be interesting for researchers and requires investigation. A more detailed LC model should then be specified, modeling local dependencies and allowing for error in the variable *rent benefit*. However, this is outside the scope of the present study. We assume this to be classification error, and therefore want this specific cell to contain zero persons. Research has previously been done regarding the relation between home ownership and marital status (Mulder, 2006). A research question here could be whether married individuals more often live in a house they own compared to non-married individuals. Therefore, a variable indicating whether a person is married or not is included in the latent class model as a covariate. The three datasets used are discussed in more detail below:

- **Registration of addresses and buildings (BAG):** A register with data on addresses containing information about its buildings, owners and inhabitants originating from

municipalities from 2013. Register information is obtained from persons who filled in the LISS studies and who declared that we are allowed to combine their survey information with registers. In total, this left us with $3,011$ individuals. From the BAG we used a variable indicating whether a person "owns" / "rents" / "other" the house he or she lives in. Because our research questions mainly relate to home-owners, we recoded this variable into "owns"/"rents or other", this variable does not contain any missing values.

- **LISS background study:** A survey on general background variables from January 2013. From this survey we also have $3,011$ individuals. We used the variable *marital status*, indicating whether someone is "married" / "separated" / "divorced" / "widowed" / "never been married". As we are only interested in whether a person is married or not, we recoded this variable in such a way that "married" and "separated" individuals are in the recoded "married" category and the "divorced", "widowed" and "never been married" individuals are in the "not married" category. It is difficult to handle a category as "separated" in such a situation. However, separated individuals are technically still married. Although they can in theory be more likely to live out of the registered address, it is difficult to make assumptions and therefore we decided to recode them into the category "married". This variable did not contain any missing values. We also used a variable indicating whether someone is a "tenant" / "sub-tenant" / "(co-) owner" / "other". We recoded this variable in such a way that we distinguish between "(co-) owner" and "(sub-) tenant or other". This variable had 14 missing values.

- **LISS housing study:** A survey on housing from June 2013. From this survey we used the variable *rent benefit*, indicating whether someone "receives rent benefit" / "the rent benefit is paid out to the lessor" / "does not receive rent benefit" / "prefers not to say". Because we are not interested in whether someone receives the rent benefit directly or indirectly, we recoded the first two categories into "receiving rent benefit". No one selected the option "prefers not to say". For this variable, we had $2,232$ missing values resulting in 779 observations. This is caused by the fact that another variable, indicating whether someone rents their house, was used as a selection variable. Dependent interviewing has been used here. Only the individuals indicating that they rent their house in this variable were asked if they receive rent benefit. This selection variable could also have been used as an indicator in our LC model. However, because of the strong relation between this variable and the rent benefit variable we decided to leave it out of the model.

Table 2.1: cross-table between the own/rent variable originating from the LISS background survey and the own/rent variable originating from the BAG register.

		register	
		rent	own
background survey	rent	902	155
	own	48	1892

These datasets are linked on person level where matching is done on person identification numbers. In addition, matching could also have been done on date, since the surveys were conducted at different time points within 2013. However, mismatches on dates are a source of classification error, and are therefore left in for illustration purposes. Although it is not necessarily the case in practice, the assumption is made that the covariate 'rent benefit' is measured without error, so we are able to apply the LC model investigated in the simulation study in practice. In Table 2.1, it can be seen that 48 individuals rent a home according to the BAG register, while stating to own a home in the LISS background survey. Furthermore, 155 individuals own a home according to the BAG register, while stating that they rent a home in the LISS background survey.

Not every individual is observed in every dataset. This causes some missing values to be introduced when the different datasets are linked on a unit level. These records are not missing, but they are considered as non-sampled individuals. Full Information Maximum Likelihood was used to handle the missing values in the indicators (Vermunt & Magidson, 2013b, pp.51-52).

Table 2.2: Entropy R^2 of the restricted conditional model; classification probabilities of the indicators and marginal probabilities of the covariates. The covariate *rent benefit* takes information of 779 individuals into account and *marital status* variable of 3011 individuals.

			output
entropy R^2			0.9380
classification probability	background survey	P(rent|LC rent)	0.9344
		P(own|LC own)	0.9992
	register	P(rent|LC rent)	0.9496
		P(own|LC own)	0.9525
P(rent benefit)			0.3004
P(married)			0.5284

The MILC method is applied to impute the latent variable *home owner/renter* by using two indicator variables and two covariates using the *restricted conditional model*. For results when the *unconditional* and the *conditional* model are applied we refer to Appendix A. In Table 2.2 classification statistics about the models are given, indicating how we can compare the results

of this model to the information we obtained in the simulation study. Both the entropy R^2 and the classification probabilities are comparable to conditions we tested in the simulation study and in which the MILC method appeared to work very well. The classification probabilities for the LISS background survey and the BAG register indicate that they both have a high quality, but are both error prone. Furthermore, P(married) and P(rent benefit) cannot be compared directly to the set up of the simulation study, but information provided by the covariates is taken into account in the entropy R^2.

Table 2.3: The first block represents the (pooled) marginal proportions of the variable *own/rent*. The second block represents the (pooled) proportions of the variable *own/rent* for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable *own/rent* for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last row represents the restricted conditional model used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	estimate	95% CI	estimate	95% CI
register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
background survey	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
restricted conditional model	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
	P(own; rent benefit)		P(rent; rent benefit)	
	estimate	95% CI	estimate	95% CI
register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
background survey	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
restricted conditional model	0.0000	-	0.2978	[0.2649; 0.3307]
	P(own; no rent benefit)		P(rent; no rent benefit)	
	estimate	95% CI	estimate	95% CI
register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
background survey	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
restricted conditional model	0.0213	[-0.0116; 0.0542]	0.6773	[0.6444;0.7102]

For the two variables measuring home ownership, we can see from the cell totals in Table 2.3 whether individuals who say to own their home also receive rent benefit, which is not allowed. However, in practice these discrepancies can be caused by the fact that people make mistakes when filling in a survey, or for example because people were moving during the period the surveys took place. Furthermore, the total number of individuals who can be found in the table of the LISS background study are only 779, and for the BAG register 772. This is because only the people indicating that they rented a house in the LISS Housing study were asked the question whether they received rent benefit. For the LISS background study we see that eight individuals are in the cell representing the impossible combination of owning a house and receiving rent benefit, and for the register four. If we investigate the cell proportions

estimated by the MILC method, we see that both the *conditional* and the *unconditional model* replicate the structure of the indicators very well, but that individuals are still assigned to the cell of the impossible combination (see Appendix A). To get this correctly estimated, we need the restricted conditional model. The marginals of the variable *own/rent* (in the upper block of Table 2.3) for the different models are all very close to each other, and closer to the estimates in the BAG register than to the estimates of the LISS background study. Also note that individuals with missing values on the variable *rent benefit* are not taken into account in the 2×2 table of *rent benefit* $\times$ *own/rent*.

Table 2.4: The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The third row represents the restricted conditional model used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the variable *owning/renting* a house.

	intercept		marriage	
	estimate	95% CI	estimate	95% CI
register	2.4661	[2.2090; 2.7233]	-1.2331	[-1.3901; -1.0760]
background survey	2.7620	[2.4896; 3.0343]	-1.3041	[-1.4678; -1.1405]
restricted conditional model	2.7712	[2.5036; 3.0389]	-1.3817	[-1.6493; -1.1140]

After we investigated the cross table between *home ownership* and *rent benefit*, we were also interested in whether *marriage* can predict *home ownership*. When we consider the BAG register, we see in Table 2.4 that the estimated odds of owning a home when not married are $e^{-1.2331} = 0.29$ times the odds when married, while they are $e^{-1.3041} = 0.27$ when the LISS background survey is used. It is interesting to see that when the *restricted conditional MILC model* is used to obtain an estimate that also corrects for the impossible combination of owning a house and rent benefit, we see that this coefficient is even a little less strong, namely $e^{-1.3817} = 0.25$. Overall, these results show us that although non-married individuals are approximately equally likely to own or rent a house, married individuals are three times more likely to own a house than to rent one.

2.5 Discussion

In this paper we introduced the MILC method, which combines latent class (LC) analysis with edit restrictions and multiple imputation (MI) to obtain estimates for variables of which we had multiple indicators in a combined dataset. We distinguished between invisibly present and visibly present errors (commonly solved by edit restrictions), and argued the need for a method that takes them into account simultaneously. We evaluated the MILC method in terms of its ability to correctly take impossible combinations and relations with other

variables into account. We assessed these relations by investigating the bias of $\hat{\theta}$, coverage of the 95% confidence interval, and $se/sd(\hat{\theta})$ in different conditions in a simulation study. The performance of MILC appeared to be mainly dependent on the entropy R^2 value of the LC model. We concluded here that a different quality of the combined dataset is required to obtain unbiased estimates and standard errors for different types of estimates. In cases of 2×2 tables including an edit restriction, a higher quality of the combined dataset was required (entropy R^2 of 0.90), while unbiased estimates and standard errors for logit coefficients can already be obtained with an entropy R^2 value of 0.60.

An example of a combined dataset containing data from the LISS panel and the BAG register were shown to have adequate entropy R^2 and we investigated the MILC method using the *unconditional model*, the *conditional model* and the *restricted conditional model*. All models can potentially be used when using the MILC method in practice. However, if there are edit restrictions within the data that need to be taken into account, only the *restricted conditional model* is appropriate. In light of our main findings, the MILC method can be seen as an alternative for methods previously used for handling visibly and invisibly present errors. This was done either separately using latent variable models and edit rules, or simultaneously by Manrique-Vallier & Reiter (2016), by using one file and one measurement.

It is important to discuss the limitations of the current simulation study. A number of limitations of the current study are related to the assumptions we made when specifying the LC model. We assumed that the observed indicators were independent of each other given a unit's score on the latent variable, which means that when a mistake is made on an indicator originating from one source, this is independent of mistakes made on indicators from other sources. For example, if multiple indicators originate from comparable surveys, there is a probability that a respondent makes the same mistake in both surveys; this assumption is then not met. There are ways to relax this assumption by extending the LC model, but we did not investigate the performance of the MILC method if this assumption is relaxed. We also assumed that the misclassification is independent of the covariates. This is also an assumption that in some cases should be relaxed, which we did not investigate as well. Furthermore, the assumption was made that the covariates are free of error. Since this assumption is often not met, ways to relax this assumption should be investigated as well as the performance of the MILC method in such cases. At last, it was assumed that all edits applied were hard edits, while sometimes soft edits are better applicable. We applied the edits by specifying which cell in the cross table between the latent variable and a covariate should have a weight of zero, while it is also possible to fix the relevant logit parameter to a very small number. In this way,

it should be possible to apply hard or soft edit restrictions. However, we did not investigate the performance of the MILC method when edits are specified in such a manner. We also did not investigate the performance of the LC model used here when some of the previously discussed assumptions is not met.

If a researcher is interested in investigating the relationship between the imputed latent variable and many other variables, all these variables should be included in the LC model as covariates. With the LC three-step approach (Bakk et al., 2016), relationships between the imputed latent variable and other variables (not incorporated in the LC model) can be investigated as well. Edit restrictions could then be added later on as well. However, this three-step approach has not been incorporated in the MILC framework. More investigation can also be done on how the MILC framework handles missing values within covariates, linkage errors and selection errors. Furthermore, the current simulation study only considers dichotomous variables. The current simulation study shows us how the method works and it gives us some indications of when the method works. This simulation was also comprehensive enough to discover the relation between the quality of the results after imputation and the entropy R^2 value of the LC model. However, it should still be investigated if this relationship holds with larger numbers of indicators, covariates and larger numbers of edit restrictions, and what the exact limitations will be here. Also situations when indicators have different numbers of categories are not yet investigated.

Another point of discussion is that we used three indicators in our LC model. In practice, it is more likely that researchers find only two indicators for an underlying true measure in their combined dataset. However, a model with two indicators is not identifiable so an additional covariate is necessary. The fact that we used three indicators might seem like a disadvantage. However, a three indicator model and a two indicator plus covariate model are Markov equivalent, which means that they yield the same set of conditional inference assumptions and an identical likelihood.

It should also be noted that MILC can be applied to indicators coming from both population registers and sample surveys. When the indicators only come from sample surveys, we can use the standard rules for pooling as defined by Rubin (1987). However, when at least one of the indicators is sourced from a complete population register, we can choose to either only impute the survey variables, and weight them to appropriately represent the population variables, or we can choose to impute both the survey and population variables, and use adjusted rules for pooling (Vink & van Buuren, 2014). We use these adjusted rules because in

the case of register indicators all sampling variability is captured by the between imputation variance, so the within variance should be left out of the equation. In this paper, we consider the situation where samples and population registers are linked on a unit level, resulting in a combined dataset consisting of only the individuals that were also in the survey sample. However, it is important to be aware of necessary adjustments when population registers are used.

PART II

Extending the latent class model of the MILC method

Chapter 3

Estimating the number of serious road injuries per vehicle type in the Netherlands using Multiple Imputation of Latent Classes

Abstract

Statistics published by official agencies are often generated using population registries, which are likely to contain classification errors and missing values. A method that simultaneously handles classification errors and missing values is Multiple Imputation of Latent Classes (MILC). In this paper, the MILC method is applied to estimate the number of serious road injuries per vehicle type in the Netherlands and to stratify the number of serious road injuries per vehicle type into relevant subgroups using data from two registries. For this specific application, the MILC method is extended to handle the large number of missing values in the stratification variable *region of accident* and to include more stratification covariates. After applying the extended MILC method, a multiply imputed dataset is generated that can be used to create statistical figures in a straightforward manner, and that incorporates uncertainty due to classification errors and missing values in the estimate of the total variance.

This chapter is accepted for publication as: L. Boeschoten, T. De Waal, J.K. Vermunt (2019) Estimating the number of serious road injuries per vehicle type in the Netherlands using Multiple Imputation of Latent Classes *Journal of the Royal Statistical Society Series A: Statistics in Society*

3.1 Introduction

When statistics are published by government or other official agencies, population registries are often utilized to generate these statistics. Here, caution is advised as population registries are collected for administrative purposes so they may not align conceptually with the target of interest. Furthermore, they are likely to contain process delivered classification errors. Another issue is that population registries are likely to not have registered every single unit in the population of interest, so the population registry is not complete.

An official agency dealing with the issues of classification errors and missing units in registries when generating statistics is the Institute for Road Safety Research (in Dutch Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, abbreviated as SWOV). An important statistic SWOV publishes every year is the number of serious road injuries in the Netherlands. The number of serious road injuries is important because it is used to define the road safety target (Reurings & Stipdonk, 2011). To gain more insight in the total number of serious road injuries, it can be further stratified by vehicle type, injury severity and region (Reurings & Bos, 2012). When estimating the number of serious road injuries in the Netherlands, SWOV uses information from police and hospital registries. These registries contain classification errors and are incomplete. SWOV estimates the number of units missing in both registries by a method based on capture-recapture (Reurings & Stipdonk, 2011). However, a procedure to correct for classification errors and missing values within the observed cases has not been applied.

A method to simultaneously deal with classification errors and missing values within the observed cases is the recently proposed method Multiple Imputation of Latent Classes (MILC; Chapter 2). The MILC method combines two existing statistical methods: multiple imputation and latent class (LC) analysis. To apply the MILC method, it is necessary to have multiple population registries that can be linked on a unit level. All registries are required to contain identifier variables for their cases which makes it possible to link the information for a specific case in one registry to its information in the other registries. In such a combined dataset, variables are selected that measure the same construct but originate from the different registries. They are used as indicators of a latent variable of which it can be said that it contains the *true* scores which are estimated using a latent class model. Information from the latent class model is then used to create multiple imputations of the *true* variable. The multiply imputed datasets can be used to generate statistics of interest, graphs or frequency tables. Uncertainty due to classification errors and missing cases is reflected in the differences

between the imputations and is incorporated in the estimate of the total variance (Rubin, 1987, p.76).

In this paper, the MILC method is applied on a linked dataset containing a police and a hospital registry, to estimate the number of serious road injuries per vehicle type. Next, two variables measuring vehicle type are used as indicators of a latent variable measuring the *true vehicle type*. Due to the way this dataset is constructed, a special feature of this dataset is that whenever one of these two indicators is missing, the other is observed. To stratify the serious road injuries into relevant groups, covariates are included in the latent class model.

A statistic that is currently not straightforward to estimate is the number of serious road injuries per vehicle type per region, because the variable *region of accident* is only observed in the police registry and contains many missing cases. To estimate this statistic, the MILC method is extended in two ways. First, the MILC method is extended to simultaneously estimate two latent variables (*vehicle type* and *region of accident*). For the latent variable *vehicle type*, two imperfectly measured indicators are specified. For the latent variable *region of accident*, one indicator (containing missing values) is assumed to be a perfect representation of the latent variable, next to a second, imperfectly measured, indicator. Second, the MILC method is extended to incorporate more covariates for investigating relevant stratifications in general. In the remainder of this paper, we refer to this as the *extended MILC method*.

In the next section, a more detailed description of the data on which the extended MILC method is applied is given. In the third section, a detailed description is given of how the extended MILC method is applied to the unit-linked police-hospital dataset. In addition, an illustrative simulation study is performed. Here, the results obtained after applying the extended MILC method are compared to results obtained after applying a more traditional hierarchical assignment procedure. In the fourth section, the output from the latent class model and the number of serious road injuries are discussed.

3.2 Background

The extended MILC method is applied on a unit linked dataset containing a police and a hospital registry. It is applied separately to datasets from 1994, 2009 and 2014 as the quality of the registries has changed substantially over time. In this section, the process of constructing these datasets is described and variables of interest are discussed in more detail.

For every year, units observed in the two sources are linked by using information on person and accident characteristics (Reurings & Stipdonk, 2009). Changes in registration systems over time influenced the success rate of the linking procedure. In addition, a weighting factor was determined for many of the individual cases (Bos et al., 2017).

3.2.1 Variables measuring *Vehicle type*

As can be seen in Table 3.1, the variable *vehicle type* is observed in both the police and the hospital registry and has nine categories. The categories make a distinction between injuries caused by motorized vehicles (with an 'M' in the category label) and non-motorized vehicles (with an 'N' in the category label). For example, there is a category 'M-bicycle' and 'N-bicycle'. The difference between these categories is that for the category 'M-bicycle', the injured person was on a bike and got into an accident with a motorized vehicle, while for the category 'N-bicycle', the injured person was on a bike and there was no motorized vehicle involved in the accident. The distinction between motorized and non-motorized is important because it provides information on the cause of the injury. For example, when the number of injuries increases in the category 'N-bicycle', it can be caused by unsafe bicycle lanes. If the number of injuries increases in the category 'M-bicycle', it can be caused by a high speed limit on roads shared by cars and bicycles.

As shown in Table 3.1, many injuries were classified differently by the police and the hospital. In addition, it can also be seen that injuries in the 'non-motorized' ('N') categories are particularly often missing in the police registry, as the police is generally not involved in, for example, one-sided bicycle accidents. Also note that the category 'N-other' is not observed in the police registry at all.

3.2.2 Variables describing relevant subgroups

Besides estimating the number of serious road injuries per vehicle type, stratifications in relevant subgroups need to be made, such as *age*, *gender*, *injury severity* or *region of accident*. To be able to make such stratifications, the variables need to be included as covariates in the latent class model that is used to estimate *true vehicle type*.

The reason for estimating the latent class model, is to create imputations for *true vehicle type* for every observed case. To be able to stratify all cases, the covariates need to be observed completely as well. For the variables *age*, *gender* and *injury severity* this is the case. For the

Table 3.1: Cross-table between the variables measuring vehicle type originating from the police registry (colums) and from the hospital registry (rows) for the years 1994, 2009 and 2013. Note that there are no observations for the category 'Non motorized - other' in the police registry. Also note that 'NA' means 'missing value'

		NA	1.	2.	3.	4.	5.	6.	7.	9.	Total
1994											
	NA	-	561	245	318	122	42	137	90	14	1,529
	1. M car	918	2,596	11	72	12	22	25	2	1	3,659
	2. M moped	702	29	1,131	21	60	2	8	2	1	1,956
	3. M bicycle	397	40	70	1,111	2	1	53	25	4	1,703
	4. M motorcycle	347	16	41	2	633	3	0	0	0	1,042
	5. M other	450	408	106	104	35	50	116	8	2	1,279
	6. M pedestrian	421	128	37	231	4	5	537	5	5	1,373
	7. N bicycle	3,625	28	41	221	3	3	11	296	3	4,231
	8. N other	34	1	0	2	0	4	0	2	0	43
	9. N pedestrian	94	2	2	2	0	0	20	6	22	148
	Total	6,988	3,809	1,684	2,084	871	132	907	436	52	16,963
2009											
	NA	-	209	111	126	38	20	62	26	6	598
	1. M car	779	969	8	29	8	17	3	0	0	1,813
	2. M moped	1,117	4	611	10	23	20	2	0	0	1,787
	3. M bicycle	565	23	17	701	0	9	20	9	0	1,344
	4. M motorcycle	668	9	74	2	367	6	0	0	0	1,126
	5. M other	350	51	40	21	11	23	23	1	1	521
	6. M pedestrian	363	39	15	62	2	2	202	2	2	689
	7. N bicycle	6,369	17	22	161	2	4	5	144	4	6,728
	8. N other	99	0	2	4	0	0	0	4	1	110
	9. N pedestrian	136	0	1	4	0	0	6	8	16	171
	Total	10,446	1,321	901	1,120	451	101	323	194	30	14,887
2013											
	NA	-	59	29	33	15	36	11	5	1	189
	1. M car	877	566	3	1	4	65	3	0	0	1,519
	2. M moped	2,220	8	419	3	167	63	2	1	0	2,883
	3. M bicycle	944	4	11	451	0	155	10	7	0	1,582
	4. M motorcycle	69	0	10	0	21	3	0	0	0	103
	5. M other	556	18	8	1	19	27	4	0	0	633
	6. M pedestrian	392	2	3	30	0	64	123	0	1	615
	7. N bicycle	7,230	12	7	41	1	29	2	44	1	7,367
	8. N other	13	0	0	0	0	0	0	0	0	13
	9. N pedestrian	117	0	0	1	0	4	2	0	5	129
	Total	12,418	669	490	561	227	446	157	57	8	15,033

variable *region of accident*, this is a problem, as this variable is only observed in the police registry.

To solve the issue of missing values in the variable *region of accident* the traditional MILC method is extended in such a way that missing values in the variable *region of accident* are

Figure 3.1: A map of the Netherlands

imputed simultaneously while the latent variable *true vehicle type* is estimated. To create these imputations, information is used from the variable *region of hospital*, which is observed for the cases that contain missing values for the variable *region of accident*. The two variables have a strong, but not perfect, relationship. For example, from the serious road injuries in 2013 of which the injured person was in a hospital in Groningen, 53 were also registered to have taken place in Groningen, while 12 of those accidents were registered to have taken place in Friesland, a neighboring region of Groningen. There was also one person in a hospital in Groningen of which the accident was registered to be in Zuid-Holland, which is quite far away from Groningen (see Figure 3.1 for the regions of the Netherlands). A reason for this observation can be classification error in one of the registries or incorrect linkage of a case in the police registry to a case in the hospital registry (wrongfully assuming the cases contained the same person). However, it is also possible that this person indeed had a road accident

in Zuid-Holland and was transfered to a hospital in Groningen because it was closer to the person's home or it could provide a form of specialized health care.

3.3 Applying the extended MILC method

In this section, it is described step-by-step how the extended MILC method is applied to estimate the number of serious road injuries per vehicle type in the Netherlands. The procedure of applying the extended MILC method starts with the dataset that is linked and processed as described in the previous section.

3.3.1 Bootstrapping for parameter uncertainty

In order to account for parameter uncertainty when applying the extended MILC method, we use a non-parametric bootstrap procedure. This involves creating m bootstrap samples by drawing observations from the observed data set with replacement. Subsequently, for each bootstrap sample, the latent class model of interest is estimated and the m imputations are created using the m sets of parameter values obtained. This is preferable over creating imputations based the maximum-likelihood estimates obtained with the observed data, which would imply ignoring the uncertainty regarding the estimated parameters of the LC model. Thus, by applying a non-parametric bootstrap procedure, parameter uncertainty is incorporated in the final pooled standard error estimates of the statistics of interest.

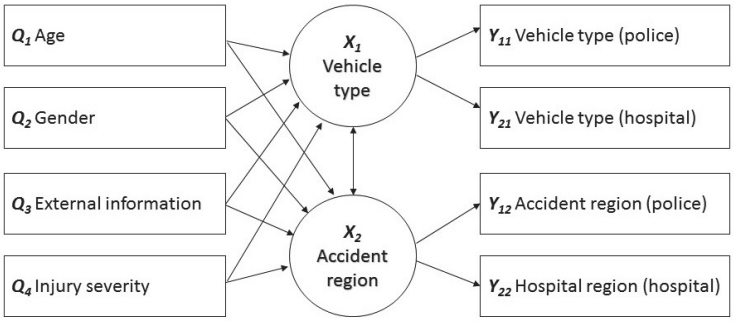

Figure 3.2: Graphical overview of the latent class model specified in Latent GOLD.

3.3.2 Specifying the latent class model

The second step of the extended MILC method is specifying the latent class model. The latent class model is estimated separately to each bootstrap sample so that the differences

Table 3.2: Cross-table between the variables region of hospital (columns) and region of accident (rows) for the years 1994, 2009 and 2013. Note that 'NA' means 'missing value'

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	Total
1994													
NA	345	419	213	627	772	499	1,152	1,140	123	997	590	111	6,988
1. Groningen	314	4	2	5	2	1	4	2	0	0	2	1	337
2. Friesland	17	393	5	7	0	1	1	5	0	3	1	2	435
3. Drenthe	7	3	230	14	1	1	1	3	0	1	0	0	311
4. Overijssel	2	3	26	711	7	4	6	4	0	9	2	4	778
5. Gelderland	2	0	3	112	977	108	10	23	1	46	4	3	1,289
6. Utrecht	3	3	1	2	52	564	38	7	2	6	3	1	682
7. N.-Holland	4	2	2	6	15	11	1,538	29	1	14	9	4	1,635
8. Z.-Holland	6	4	7	8	16	22	30	1,564	4	20	8	2	1,691
9. Zeeland	0	0	0	0	2	0	1	9	212	24	0	0	248
10. N.-Brabant	1	2	1	5	60	6	17	35	2	1,550	33	0	1,712
11. Limburg	0	2	1	1	19	2	5	3	1	12	690	1	737
12. Flevoland	0	1	0	6	6	5	10	3	0	0	0	89	120
Total	751	836	491	1,504	1,929	1,224	2,813	2,827	3,46	2,682	1,342	218	16,963
2009													
NA	435	586	267	667	1,523	865	2,014	1,728	151	1,185	840	185	10,446
1. Groningen	186	5	2	2	3	0	3	1	0	4	1	0	207
2. Friesland	23	200	3	3	0	1	1	0	0	0	0	0	231
3. Drenthe	48	0	91	16	1	0	1	3	1	4	2	0	167
4. Overijssel	2	2	3	265	2	0	2	5	0	1	1	0	283
5. Gelderland	1	2	0	51	516	58	5	5	0	20	2	0	660
6. Utrecht	1	2	0	3	26	323	23	1	1	2	0	0	382
7. N.-Holland	0	3	2	1	10	11	673	11	2	6	12	0	731
8. Z.-Holland	2	3	1	3	6	19	13	683	0	6	4	0	740
9. Zeeland	0	0	0	0	0	0	2	3	80	8	1	0	94
10. N.-Brabant	1	0	0	0	23	4	9	14	0	491	6	0	548
11. Limburg	0	0	0	1	7	0	4	1	1	3	300	1	318
12. Flevoland	1	13	0	22	5	3	6	3	0	0	0	27	80
Total	700	816	369	1,034	2,122	1,284	2,756	2,458	236	1,730	1,169	213	14,887
2013													
NA	392	534	372	857	1,696	870	2,643	2,286	324	1,475	815	154	12,418
1. Groningen	53	0	0	0	0	0	0	1	1	2	1	0	58
2. Friesland	12	77	0	1	0	1	1	0	0	0	0	0	92
3. Drenthe	18	0	36	8	1	1	0	0	0	0	1	0	65
4. Overijssel	0	0	0	180	1	0	0	2	0	0	2	0	185
5. Gelderland	0	0	0	37	313	30	2	2	0	5	2	0	391
6. Utrecht	0	0	0	0	13	178	15	1	1	3	2	0	213
7. N.-Holland	2	0	0	1	3	7	492	3	0	0	1	1	510
8. Z.-Holland	1	0	1	2	4	8	14	439	1	6	1	2	479
9. Zeeland	0	0	0	0	0	0	0	8	57	5	0	0	70
10. N.-Brabant	1	1	2	0	19	1	3	11	0	293	3	0	334
11. Limburg	0	0	0	2	14	0	0	3	1	3	141	0	164
12. Flevoland	1	2	0	10	2	2	15	0	0	0	0	22	54
Total	480	614	411	1,098	2,066	1,098	3,185	2,756	385	1,792	969	179	15,033

between the parameters in the different latent class models reflect parameter uncertainty. A graphical overview of the specified latent class model can be found in Figure 3.2. First, the latent variable measuring *vehicle type* (X_1) is specified. The variables measuring vehicle type originating from the police registry (Y_{11}) and from the hospital registry (Y_{21}) are specified as indicators of this latent variable. Note that this notation differs from traditional notation where X variables are predictors and Y variables are responses, e.g. in regression analysis. As was discussed in Section 3.2, the vehicle type indicator variables contain nine categories in total, six representing motorized vehicles and three representing non-motorized vehicles. However, specifying nine latent classes would be problematic, since the number of observed non-motorized accidents in the police registry is very low. Therefore, the non-motorized categories are grouped into one category resulting in the specification of a seven class model. By saving the original scores of this indicator variable in separate variables, these can be re-assigned to the accidents which were assigned to the latent class 'accidents without motorized vehicle' after multiple imputation. For this, the proportions of the categories in the observed data are used.

Second, all covariates of interest need to be included in the latent class model, because otherwise point estimates describing the relationship between a latent variable and an excluded covariate will be biased (Bolck et al., 2004). As discussed in Section 3.2, the variable *region of accident* cannot be included directly as a covariate as it contains a large proportion of missing values. Therefore, multiple imputations are created for this variable to be able to stratify for the variable *vehicle type* over the different regions in the Netherlands. For this purpose, a second latent variable is specified to measure *region of accident* (X_2). The first indicator is the variable *region of accident* measured in the police registry (Y_{12}). The second indicator variable is *region of hospital* (Y_{22}). Since the first indicator is actually the variable for which imputations are created, the relationship between the latent variable and the indicator variable is restricted such that if the indicator variable is observed, this score is assigned directly to the latent variable as well. Only if this indicator variable contains a missing value, the outcomes of this latent class model are used.

Other covariates needed to make relevant stratifications can be included in the latent class model directly, since they do not contain any missing values. The other covariates included in the latent class model are:

- Age: $0-17$; $18-44$; $46-69$; $70+$ (Q_1).
- Gender: Male; Female (Q_2).

- External information: Standard; Falling; Non-public road; No driving vehicle; Other (Q_3).
- Injury severity using Abbreviated Injury Scale (AIS), an anatomical scoring system where injuries are ranked on a scale from one to six. As 'one' represents 'minor injuries' and 'six' represents 'unsurvivable injuries', these do not fit in the scope of this research, as this research pertains to 'serious road injuries'. Therefore, the following scores on AIS are included: 'two' means 'Moderate'; 'three' means 'Serious'; 'four' means 'Severe'; 'five' means 'Critical' (Wong, 2011) (Q_4).

To ensure that all parameters can be estimated for each bootstrap sample, only main effects of the covariates are included in the latent class model.

The latent class model for response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})$ is:

$$\begin{aligned}\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q}) = \sum_{x_1=1}^{7} \sum_{x_2=1}^{12} \prod_{l_1=1}^{2} \mathrm{P}(Y_{l_1,1} = y_{l_1,1} \mid X_1 = x_1) \cdot \\ \prod_{l_2=1}^{2} \mathrm{P}(Y_{l_2,2} = y_{l_2,2} \mid X_2 = x_2) \cdot \\ \mathrm{P}(X_1 = x_1, X_2 = x_2 \mid \mathbf{Q} = \mathbf{q}).\end{aligned} \tag{3.1}$$

In this latent class model, X_1 represents the latent variable *vehicle type* with seven classes and X_2 represents the latent variable *region of accident* with 12 classes. Furthermore, $\mathbf{Q}$ represents the covariate variables and $\mathbf{Y}$ represents the indicator variables, where l_1 stands for the two indicator variables corresponding to X_1 and l_2 for the two indicator variables corresponding to X_2 (which corresponds to what can be seen in Figure 3.2). The latent class model is estimated using Latent GOLD 5.1 (Vermunt & Magidson, 2013a), where the recommendations by Vermunt et al. (2008) for large datasets have been followed to ensure convergence. See Appendix B for the Latent GOLD syntax used.

By specifying the previously described latent class model, the first assumption made is that the probability of obtaining a specific response pattern is a weighted average of all conditional response probabilities, also known as the mixture assumption. Second, the assumption is made that the observed indicators are independent of each other given a unit's score on the underlying true measure. In other words, this means that if a classification error is made in the police registry, we assume that this is independent of the probability of also having a classification error in the hospital registry. For most cases this assumption can be considered

realistic, since the police registry and the hospital registry are generally filled out by two different and independent persons. In rare situations, dependencies might arise. For example, in a 'hit-and-run' situation, both registries will probably be filled out based on information provided by the victim and are therefore not independent. Third, the assumption is made that the misclassification in the indicators is independent of the covariates. It is unlikely that scores on covariates such as *age* or *gender* will influence this. However, for example for the variable *external information*, it can be the case that if an accident takes place outside the public road, it is more difficult for the police to reach this location and therefore the probability of an error can increase. Fourth, the assumption is made that the covariate variables are free of error. This is, of course, an unrealistic assumption, especially given the substantial amounts of classification error found in the *vehicle type* indicator variables. At this point we unfortunately do not have any information available about the extent of possible classification errors in the other variables. However, these errors are considered less problematic as long as they are random. Lastly, assumptions are made with respect to the missingness mechanisms present in the data. More specifically, the mechanism that governs the probability each data point has of being missing is considered Missing At Random (MAR) for the variables *vehicle type* observed in the police registry (Y_{11}) and *region of accident* (Y_{12}), as the probability of being missing is larger for 'non-motorized' vehicles, which is measured by the hospital registry (Y_{21}). Formally, it can be stated that Y_{11} consists of a part $Y_{11,\text{obs}}$ and $Y_{11,\text{mis}}$ and that a vector R_{11} can be defined

$$R_{11} = 0 \text{ if } Y_{11,\text{obs}};$$
$$R_{11} = 1 \text{ if } Y_{11,\text{mis}}.$$

As we assume the missingness mechanism to be MAR, the distribution of missing values is related to Y_{21}:

$$\text{P}(R_{11} = 0|Y_{11,\text{obs}}, Y_{11,\text{mis}}, Y_{21}) = \text{P}(R_{11} = 0|Y_{11,\text{obs}}, Y_{21}).$$

If a value is missing in Y_{11}, it is by definition missing in Y_{12} as well, as unit missingness is considered here and both variables originate from the same dataset. Furthermore, the mechanism that governs the probability of being missing is considered Missing Completely At Random for the variable *vehicle type* observed in the hospital registry (Y_{21}). Here,

$$\text{P}(R = 0|Y_{21,\text{obs}}, Y_{21,\text{mis}}, Y_{11,\text{obs}}) = \text{P}(R = 0).$$

The distribution of missing values in Y_{11} and Y_{12} is related to Y_{21}, which in itself also contains missing values. Generally this would mean that we are not dealing a MAR mechanism for

Y_{11} and Y_{12}. However, due to the special structure of our dataset in which Y_{11} and Y_{12} never contain missing values if Y_{21} contains missing values and vice versa, we are still dealing with a MAR mechanism. Cases containing missing values on all above mentioned variables are by definition not included in the dataset, and are treated separately.

The latent class model gives different forms of relevant output. The first form of relevant output is the entropy R^2. Entropy can be formally defined as:

$$\text{EN}(\alpha) = -\sum_{j=1}^{N}\sum_{x=1}^{X}\alpha_{jx}\text{log}\alpha_{jx},$$

where α_{jx} is the probability that observation j is a member of class x, X the number of classes, and N is the number of units in the combined dataset. Rescaled to values between zero and one, entropy R^2 is measured by:

$$R^2 = 1 - \frac{\text{EN}(\alpha)}{N\text{log}X},$$

where one means perfect prediction (Dias & Vermunt, 2008). In Chapter 2, it is shown that the performance of MILC method is closely related to the entropy R^2 of the corresponding latent class model.

A second form of relevant output are the conditional response probabilities. They provide us the probability of obtaining a specific response on the indicator conditional on belonging to a certain latent class. These values can be used to investigate the relationships between the indicator variables and the latent variables into detail. For example, they show us the probability of having the score 'M-car' on the indicator originating from the police registry given that the model assigned a case to the latent class 'M-car', but also the probability of having the score 'M-bicycle' on the indicator given that the model assigned a case to the latent class 'M-car'. Here, the former should be much higher compared to the latter. By comparing the conditional response probabilities to the cross-table between the variables measuring vehicle type originating from the police registry and the hospital registry (as seen in Table 3.1), it can be investigated if the latent classes identified as certain categories of *vehicle type* are related to other categories in the indicator variables in a comparable way as in the observed data. In this way, it is checked if the latent class model reflects the main relations found in the observed data, which is an important indication of adequate imputations in the next step.

Third, the posterior membership probabilities represent the probability that a unit belongs to a latent class given its combination of scores on the indicators and covariates used in the latent

class model. These values are used to create multiple imputations for the latent variables, and the exact procedure for this is described in the next section.

3.3.3 Creating multiple imputations

The posterior membership probabilities are used to create multiple imputations of the latent variables containing the *true* scores. The posterior membership probabilities can be estimated by applying Bayes' rule to the latent class model described in Equation 3.1:

$$\mathrm{P}(X_1 = x_1, X_2 = x_2 \mid \mathbf{Y} = \mathbf{y}, \mathbf{Q} = \mathbf{q}) = \frac{\mathrm{P}(X_1 = x_1, X_2 = x_2, \mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})}{\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})},$$

where

$$\begin{aligned}\mathrm{P}(X_1 = x_1, X_2 = x_2, \mathbf{Y} = \mathbf{y} | \mathbf{Q} = \mathbf{q}) = &\prod_{l_1=1}^{2} \mathrm{P}(Y_{l_1,1} = y_{l_1,1} \mid X_1 = x_1) \cdot \\ &\prod_{l_2=1}^{2} \mathrm{P}(Y_{l_2,2} = y_{l_2,2} \mid X_2 = x_2) \cdot \\ &\mathrm{P}(X_1 = x_1, X_2 = x_2 \mid \mathbf{Q} = \mathbf{q}).\end{aligned}$$

and $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})$ is defined in Equation 3.1.

Since there are two latent variables specified in this model, the joint posterior membership probabilities are obtained which represent the probability that a unit is a member of a specific latent class in the latent variable *vehicle type*, and a member of a specific latent class in the latent variable *accident region*. Since the variable *vehicle type* has seven classes and the variable *accident region* has 12 classes, there are 84 posterior membership probabilities which add up to one, and there is a different set of posterior membership probabilities for each combination of scores on the indicators and covariates. Parameter estimation was constrained in such a way that if a case had an observed score on the variable *accident region* in the police registry, this score is automatically assigned to the latent variable as well. In those cases, there are only seven posterior membership probabilities with a value larger than zero (those representing the different classes for *vehicle type* in combination with that specific region); all other posterior membership probabilities are exactly zero.

For each case in the original dataset, the posterior membership probabilities corresponding to its combination of scores on the indicators and covariates are used as a multinomial distribution to draw a joint score on both latent variables. These joint scores are then used

to create separate imputations for the variables *vehicle type* and *accident region*.

By drawing multiple times from the posterior membership probabilities, multiple imputations for both latent variables are created. The scores assigned to the latent variables can be different for the different imputations. The differences between them reflect the uncertainty due to the missing and conflicting values in the indicator variables. In Chapter 2, it was concluded that a low number of imputations, such as five is already sufficient for a correct estimation of the standard errors. However, in that simulation study the number of classes was much lower compared to the number of classes needed for this dataset. To evaluate what the appropriate number of imputations would be, the number of imputations was gradually increased and the fraction of missing information was compared between the differing numbers of imputations (J. W. Graham et al., 2007), resulting in 20 imputations. This is in line with the recommendations by Wang et al. (2005).

3.3.4 Pooling of the results

At this point, 20 imputations are created for *vehicle type* and *region of accident* for every unit in the combined dataset. The goal is to obtain estimates of interest using these imputed variables. This is done by obtaining the estimate of interest for every imputed variable, and pooling these estimates using the pooling rules defined by Rubin (Rubin, 1987, p.76). Although our context differs from the traditional statistical context for which the pooling rules were originally developed, the rules are considered appropriate for the context of multiple imputation for measurement error (Reiter & Raghunathan, 2007). For this specific research, the main estimates of interest are frequency tables.

The first step is to calculate a pooled frequency table. In other words, we take the average over the imputations for every cell in the frequency table. This can be for the imputed variable *vehicle type*, for the imputed variable *region of accident* or for a cross-table between (one of) these variables and covariate(s). A pooled cell count is obtained by

$$\hat{\theta}_j = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_{ij},$$

where θ refers to a cell count, j refers to a specific cell in the frequency table, i refers to one imputation and m refers to the total number of imputations.

Next, an estimate of the uncertainty around these frequencies is of interest. Therefore, the

pooled frequencies need to be transformed into pooled proportions:

$$\hat{\mathrm{p}}_j = \frac{\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{ij}}{\sum_{j=1}^{s}\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{ij}},$$

where s refers to the number of cells in the frequency table.

Since we work with a multiply imputed dataset, an estimate of the variance is obtained that is a combination of sampling uncertainty and uncertainty due to missing and conflicting values in the dataset. This is the total variance that consists of a 'within imputation' and 'between imputation' component:

$$\mathrm{VAR}_{\mathrm{total}_j} = \overline{\mathrm{VAR}}_{\mathrm{within}_j} + \mathrm{VAR}_{\mathrm{between}_j} + \frac{\mathrm{VAR}_{\mathrm{between}_j}}{m}.$$

$\overline{\mathrm{VAR}}_{\mathrm{within}_j}$ is the within imputation variance of $\hat{\mathrm{p}}_j$ calculated by

$$\overline{\mathrm{VAR}}_{\mathrm{within}_j} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{VAR}_{\mathrm{within}_{ij}},$$

where $\mathrm{VAR}_{\mathrm{within}_{ij}}$ is estimated as the variance of $\hat{\mathrm{p}}_{ij}$:

$$\frac{\hat{\mathrm{p}}_{ij}(1-\hat{\mathrm{p}}_{ij})}{N},$$

where N is the total size of the observed dataset and $\hat{\mathrm{p}}_{ij}$ is estimated as

$$\hat{\mathrm{p}}_{ij} = \frac{\hat{\theta}_{ij}}{\sum_{j=1}^{s}\hat{\theta}_{ij}}.$$

$\mathrm{VAR}_{\mathrm{between}_j}$ is calculated by

$$\mathrm{VAR}_{\mathrm{between}_j} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\mathrm{p}}_{ij} - \hat{\mathrm{p}_j})(\hat{\mathrm{p}}_{ij} - \hat{\mathrm{p}_j})'.$$

When $\mathrm{VAR}_{\mathrm{total}_j}$ is estimated, it can be used to estimate the standard error of $\hat{\mathrm{p}}_j$

$$\mathrm{SE}(\hat{\mathrm{p}}_j) = \sqrt{\mathrm{VAR}_{\mathrm{total}_j}}.$$

From here, the confidence interval around $\hat{\mathrm{p}}_j$ can be estimated by

$$\hat{\mathrm{p}}_j \pm Z_{0.975} \times \mathrm{SE}(\mathrm{P}_j),$$

where 0.975 corresponds to the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution for $\alpha = 0.05$. The values obtained here can simply be multiplied by N to obtain the 95% confidence intervals around the observed frequencies $\hat{\theta}_j$. Note that a standard normal distribution is assumed so problems can be encountered when dealing with very small proportions.

3.3.5 Performance of the MILC method

Chapter 2 introduced the MILC method, and evaluated the method under a range of conditions in terms of data quality. In addition, Chapter 4 extended the method for situations with longitudinal data and Chapter 5 extended the method in such a way that covariates can be included at later time-points. All research performed on the MILC method so far showed a strong relationship between the performance of the method and the entropy R^2 value of the latent class model. To investigate how the MILC method performs in comparison to the hierarchical assignment procedure traditionally used by SWOV (Bos et al., 2017), an illustrative simulation study is performed. In the theoretical population used for this

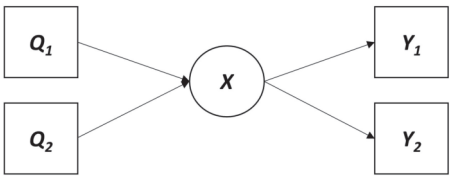

Figure 3.3: Graphical overview of the latent class model used for the simulation study

simulation study, latent variable X has two categories with probabilities

$$\begin{matrix} X = 1 \\ X = 2 \end{matrix} \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}.$$

The probability distribution of $\mathrm{P}(X, Q_1)$ is

$$\begin{matrix} & Q_1 = 1 & Q_1 = 2 \\ X = 1 & 0.48 & 0.12 \\ X = 2 & 0.32 & 0.08 \end{matrix}$$

and the probability distribution of $\mathrm{P}(X, Q_2)$ is

$$\begin{array}{c} \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Q_2=1 & Q_2=2 & Q_2=3 \end{array} \\ \begin{pmatrix} 0.36 & 0.18 & 0.06 \\ 0.24 & 0.12 & 0.04 \end{pmatrix} \end{array}$$

From this population structure, $1,000$ samples are drawn. In each sample, indicator (Y_1) of X is created with 5% misclassification and a Missing At Random (MAR) mechanism, where the probability of being missing is related to a person's score on the Q_2 covariate.

$$Q_2 = 1, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.20;$$
$$Q_2 = 2, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.15;$$
$$Q_2 = 3, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.10.$$

A second indicator (Y_2) of X is created with 15% misclassification and 5% missing cases which are Missing Completely At Random (MCAR). The latent class models had an entropy R^2 value of approximately 0.75.

The MILC method as described in Section 3.3.1 to Section 3.3.3 is applied on the sample datasets, where five bootstrap samples are drawn and subsequently five imputations of X are created. As an illustration, the MILC method is also applied without the bootstrap procedure; so with one latent class model directly estimated on the observed data and five imputations drawn from one single set of posterior membership probabilities. Furthermore, the hierarchical assignment procedure as used by SWOV is also applied. At SWOV, the score observed in the police registry (Y_1) is assigned if it is observed. Otherwise, the score observed in the hospital registry (Y_2) is assigned.

The imputations are evaluated in terms of bias, coverage of the 95% confidence interval, confidence interval width, average standard error of the estimates divided by the standard deviation over the estimates and the root mean squared error. Furthermore, the proportion of correctly classified cases is evaluated for imputation and hierarchical assignment.

To evaluate the methods, the marginals of the imputed latent variable (W) are compared to the hierarchically assigned variable (W_{ass}). In addition the estimated relationships of the latent variable with covariates ($W \times Q_1$, $W_{\text{ass}} \times Q_1$, $W \times Q_2$ and $W_{\text{ass}} \times Q_2$) are examined.

In Table 3.3 the results of the simulation study comparing the MILC method (with and

Table 3.3: Results of a simulation study where the hierarchical assignment procedure is compared to the MILC method, which is performed with and without a non-parametric bootstrap. Results are shown for the imputed mixture variable, denoted by W, and of the relationship of W with covariates Q_1 and Q_2. Results are given in terms of bias, coverage of the 95% confidence interval, confidence interval width, the average standard error of the estimate divided by the standard deviation over the estimates, and the root mean squared error.

	Bias	Coverage	CI width	se/sd	RMSE
Hierarchical assignment					
$W_{\text{ass}} = 1$	-0.0134	0.2180	0.0193	0.9981	0.0143
$W_{\text{ass}} = 2$	0.0134	0.2180	0.0193	0.9981	0.0143
$W_{\text{ass}} = 1 \times Q_1 = 1$	-0.0106	0.4220	0.0196	0.9894	0.0117
$W_{\text{ass}} = 2 \times Q_1 = 1$	-0.0028	0.8380	0.0126	0.9964	0.0043
$W_{\text{ass}} = 1 \times Q_1 = 2$	0.0107	0.3590	0.0184	0.9963	0.0117
$W_{\text{ass}} = 2 \times Q_1 = 2$	0.0027	0.8380	0.0108	1.0191	0.0038
$W_{\text{ass}} = 1 \times Q_2 = 1$	0.0012	0.3560	0.0134	0.9433	0.0052
$W_{\text{ass}} = 2 \times Q_2 = 1$	-0.1676	0.6390	0.0107	1.0053	0.1677
$W_{\text{ass}} = 1 \times Q_2 = 2$	-0.2910	0.7770	0.0066	0.9898	0.2910
$W_{\text{ass}} = 2 \times Q_2 = 2$	-0.1115	0.3050	0.0121	0.9702	0.1116
$W_{\text{ass}} = 1 \times Q_2 = 3$	-0.2261	0.5920	0.0092	1.0201	0.2261
$W_{\text{ass}} = 2 \times Q_2 = 3$	-0.3022	0.7990	0.0056	1.0552	0.3022
MILC method, bootstrap excluded					
$W = 1$	-0.0317	0.1300	0.0216	0.1425	0.042
$W = 2$	0.0317	0.1300	0.0216	0.1425	0.042
$W = 1 \times Q_1 = 1$	-0.0252	0.1660	0.0213	0.1751	0.0335
$W = 2 \times Q_1 = 1$	-0.0066	0.4410	0.0132	0.3912	0.0093
$W = 1 \times Q_1 = 2$	0.0253	0.1660	0.0205	0.1683	0.0336
$W = 2 \times Q_1 = 2$	0.0064	0.3980	0.0118	0.3628	0.0089
$W = 1 \times Q_2 = 1$	-0.0191	0.2270	0.0201	0.2151	0.0257
$W = 2 \times Q_2 = 1$	-0.0095	0.3470	0.0157	0.3278	0.0131
$W = 1 \times Q_2 = 2$	-0.0031	0.5820	0.0096	0.5341	0.0048
$W = 2 \times Q_2 = 2$	0.0191	0.1980	0.0188	0.2029	0.0255
$W = 1 \times Q_2 = 3$	0.0095	0.3150	0.0142	0.2980	0.0131
$W = 2 \times Q_2 = 3$	0.0031	0.5510	0.0085	0.4962	0.0046
MILC method including bootstrap					
$W = 1$	-0.0304	0.8880	0.1790	1.5797	0.0420
$W = 2$	0.0304	0.8880	0.1790	1.5797	0.0420
$W = 1 \times Q_1 = 1$	-0.0241	0.8950	0.1439	1.5811	0.0335
$W = 2 \times Q_1 = 1$	-0.0063	0.9050	0.0383	1.4324	0.0093
$W = 1 \times Q_1 = 2$	0.0243	0.8940	0.1437	1.5744	0.0336
$W = 2 \times Q_1 = 2$	0.0062	0.9160	0.0378	1.4887	0.0089
$W = 1 \times Q_2 = 1$	-0.0183	0.8880	0.1087	1.5375	0.0257
$W = 2 \times Q_2 = 1$	-0.0091	0.9020	0.0560	1.5192	0.0131
$W = 1 \times Q_2 = 2$	-0.0030	0.9290	0.0205	1.4125	0.0048
$W = 2 \times Q_2 = 2$	0.0183	0.8910	0.1085	1.5562	0.0255
$W = 1 \times Q_2 = 3$	0.0092	0.9050	0.0555	1.5085	0.0131
$W = 2 \times Q_2 = 3$	0.0030	0.9280	0.0200	1.4670	0.0046

without bootstrap) and the hierarchical assignment procedure are shown. We first discuss the performance of the MILC method in comparison to the hierarchical assignment method.

The results obtained with hierarchical assignment especially show substantial amounts of bias for $W_{\text{ass}} \times Q_2$ as compared to both implementations of the MILC method. For the unbiased parameters obtained when applying hierarchical assignment, the RMSE is in general lower and more stable compared to the RMSE of MILC. The fact that with hierarchical assignment, bias is especially found in the results relating to Q_2, can be explained by the fact that the missingness mechanism of Y_1 is defined by Q_2.

Comparison of the MILC method with and without bootstrap shows clearly that standard errors are very much underestimated when no bootstraps are performed; that is, coverage rates are too low and the ratios between the average standard error and the standard deviation across replications are far below one. In contrast, these ratios are larger than one when the bootstrap is included in the MILC method, meaning the standard errors are somewhat overestimated. The large difference between the two approaches is caused by the fact that the statistics we are interested in are tables containing the latent variable X. By not applying the bootstrap, one seriously underestimates the uncertainty about the latent class proportions. The fact that the bootstrap procedure yields slightly too large standard errors can be considered to be less problematic than having (much) too small standard errors.

The percentage of incorrectly classified cases is 4.5% for $X = 1$ and 10.1% for $X = 2$ when hierarchical assignment is applied (these results are not shown in Table 3.3) When the MILC method (including bootstrap) is applied, the percentage of incorrectly classified cases is 8.6% for $X = 1$ and 20.5% for $X = 2$. With hierarchical assignment, the score on one indicator variable is used per case, and the misclassification corresponds to the misclassification specified in these variables. When MILC is applied, two indicator variables are used to generate the variables under evaluation here. When assigning scores, maintaining the relationships with other variables is apparently considered more important than correctly classifying individual cases. Including interaction terms in the latent class model may possibly lead to more accurate results for the MILC method. Whether this really is the case remains to be examined, though.

3.4 Results

First, results in terms of relevant model output will be discussed. Second, substantial results obtained after creating multiple imputations for the latent variables are given.

Table 3.4: Entropy R^2 values for the latent variables *vehicle type* and *region of accident* for the years 1994, 2009 and 2013.

	Vehicle type	Region of accident
1994	0.8219	0.9050
2009	0.7444	0.8267
2013	0.8031	0.8077

3.4.1 Latent class model output

The first relevant model output from the latent class models comes in terms of the entropy R^2. A separate entropy R^2 value is estimated for the two latent variables and for each year. The results are shown in Table 3.4. These results are obtained after applying an latent class model on the original dataset. Here it can be seen that the entropy R^2 value in 2013 increased compared to 2009 for the variable *vehicle type*. Pankowska et al. (2017a) showed in their simulation studies that when a latent class model is used to correct for misclassification in combined datasets, the model also treats inconsistencies due to incorrect linkage as misclassification and thereby corrects for it in a similar way. This implies that the increase in terms of entropy R^2 in 2013 in comparison to 2009 for the latent variable *vehicle type* makes sense as the police improved their registration system in 2013. This improvement caused an increase in the number of correctly linked cases and therefore also improved the entropy R^2. The higher entropy R^2 values found for 1994 are likely to be caused by the fact that registration was performed more carefully and thorough by the police at that period, which also resulted in the lower amount missing values, as can be seen in Table 3.1.

In Table 3.5, the probability of correct classification for the indicators of both latent variables are shown, for the three different time-points, obtained after applying a latent class model to the original dataset. Class-specific response probabilities indicate the probability of having a score on the indicator variable that is equal to the latent class. A high probability of correct classification indicates that when a specific case belongs to a certain latent class, the probability is large that this same score was obtained on an indicator variable. For example, the probability of correct classification of the 1994 indicator variable 'Hospital' for the latent class *vehicle type* = M car is 0.8226. This means that the probability of having scored 'M car' on the indicator variable *vehicle type* measured by hospital is 0.8226 given that this case truly belongs to the latent class 'M car'.

When looking at the probabilities of correct classification for a specific latent class, the two probabilities corresponding to the two indicators are often not equal. This may be due to

Table 3.5: Class-specific response probabilities for latent variables 'vehicle type' and 'region of accident' for the years 1994, 2009 and 2013.

	1994		2009		2013	
Vehicle type						
	Hospital	Police	Hospital	Police	Hospital	Police
1. M car	0.8226	0.9782	0.8004	0.9742	0.9590	0.8973
2. M moped	0.8458	0.9781	0.7194	0.9786	0.9693	0.8848
3. M bicycle	0.7393	0.9170	0.7635	0.9620	0.9263	0.7376
4. M motorcycle	0.8353	0.9686	0.8876	0.9129	0.0774	0.7577
5. M other	0.6890	0.0578	0.5276	0.2629	0.0000	0.4243
6. M pedestrian	0.7132	0.8213	0.8758	0.8104	0.5358	0.6412
7. N all	0.9920	0.6162	0.9916	0.5273	0.9931	0.3897
Region of accident						
	Region of hospital	Region of accident	Region of hospital	Region of accident	Region of hospital	Region of accident
1. Groningen	0.9351	1	0.8798	1	0.9167	1
2. Friesland	0.9063	1	0.8740	1	0.8433	1
3. Drenthe	0.7338	1	0.5897	1	0.6556	1
4. Overijssel	0.9103	1	0.9290	1	0.9675	1
5. Gelderland	0.7551	1	0.7961	1	0.8119	1
6. Utrecht	0.8292	1	0.8259	1	0.8149	1
7. N.-Holland	0.9378	1	0.9267	1	0.9673	1
8. Z.-Holland	0.9240	1	0.9248	1	0.9094	1
9. Zeeland	0.8506	1	0.8248	1	0.7941	1
10. N.-Brabant	0.9084	1	0.9055	1	0.8884	1
11. Limburg	0.9397	1	0.9466	1	0.8725	1
12. Flevoland	0.7771	1	0.5374	1	0.4694	1

differences in the quality of the data. A low probability of correct classification can be caused by the fact that for this specific latent class, this category is observed many times in one indicator (here this is often the indicator originating from the hospital registry), while in the other indicator (originating from the police registry), these cases are often missing. This can clearly be seen for the latent class 'N-all'. Conditional on truly belonging in this latent class, the probability of obtaining this score on the hospital indicator was 0.9920 in 1994. In other words, almost everyone who is assigned to this class by the model, obtained this score in the hospital registry as well. However, the probability of obtaining this score by the police is only 0.6162. A substantial part of the cases belonging to this latent class obtained another score or no score at all by the police.

In general, it can be seen that the probabilities of correct classification for the police indicator in 1994 and 2009 are larger compared to the hospital indicator for all motorized classes except the class 'motorized, other' and the 'all non-motorized' category. However, in 2013 all probabilities of correct classification are higher for the hospital indicator compared to the

police indicator. This result might be related to the improvement in the linking in 2013. An exception is the category 'M motorcycle', which is the only category with a probability of correct classification below 0.90 in the hospital registry. This is caused by the fact that some of the hospitals used a different registration system, that categorizes both motorcycles and mopeds into the motorcycle category.

When investigating the probabilities of correct classification for the latent variable *region of accident*, it can be seen that they are all exactly one for the indicator variable *region of accident*. Conditional on being in a specific class in the latent variable *region of accident*, the probability of obtaining the same score on the indicator variable *region of accident* is exactly one. This restriction was imposed on the latent class model. The probabilities of correct classification of the indicator variable *region of hospital* now show us the probability that conditional on an accident truly happening in a specific region, the probability of also going to a hospital in that same region. These probabilities are generally quite high and stable over the different time-points. The regions Drenthe and Flevoland stand out because the probability of going to a hospital in these regions when having a serious road accident in this region is somewhat lower compared to other regions.

3.4.2 Pooled results output

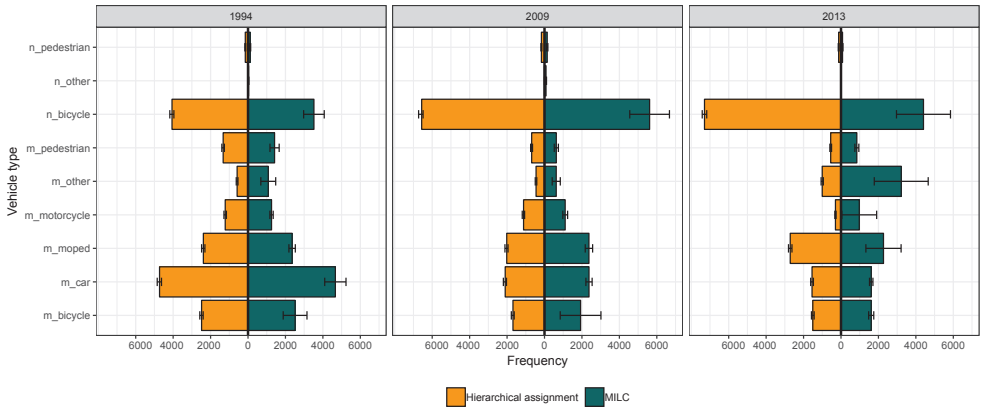

Figure 3.4: The three graphs represent results obtained for three different years. On the left side of each graph, the number of serious road injuries per vehicle type and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied. On the right side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied.

In Figure 3.4, the number of serious road injuries per vehicle type are shown for the three different years investigated. For every year, the results obtained after applying the hierarchical

assignment procedure* are compared to results obtained when the extended MILC method is applied. Here, it can be seen that in general the frequencies obtained after applying the extended MILC method are quite similar compared to the results obtained after applying the hierarchical assignment procedure. When the extended MILC method is applied, the number of cases assigned to the category 'M-other' is larger while the number of cases assigned to the category 'N-bicycle' is smaller compared to the hierarchical assignment procedure, particularly in 2013. This corresponds to a large amount of missing cases for 'N-bicycle' and a substantial amount of cases differently categorized by the police and hospital. Furthermore, in 2013 the number of cases categorized as 'M-other' by the hospital increased, while this category was often classified differently by the police (see Table 1). At last, it can be seen that the width of the 95% confidence intervals are substantially larger for all categories when the extended MILC method is applied. This directly results from the misclassification between the hospital and the police registry. Because of this misclassification, the latent class model is less certain about which value to assign to a specific case, resulting in differences between imputations and a larger estimate of the total variance. Note also that hierarchical assignment assumes that values observed in the the police register are error-free. Since this assumption is unlikely to be correct, uncertainty in the hierarchical assignment procedure is underestimated.

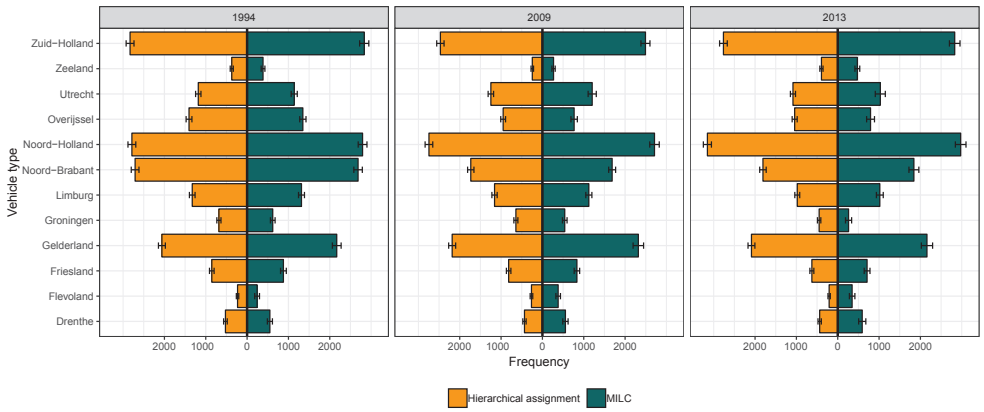

Figure 3.5: The three graphs represent results obtained for three different years. On the left side of each graph, frequencies of serious road injuries per region and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied. On the right side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied.

In Figure 3.5, the number of serious road injuries per region are shown for the three different years investigated. For every year, the results obtained after applying the hierarchical

*Note that the results presented in this paper using the hierarchical assignment procedure are not necessarily exactly equal to official statistics produced by SWOV.

assignment procedure are compared to results obtained when the extended MILC method is applied, which are very similar. The 95% confidence intervals are larger when the extended MILC method was applied compared to the hierarchical assignment procedure, but the difference is not as substantial as was the case for the variable 'vehicle type' in Figure 3.4.

3.5 Discussion

In this article, an extension of the MILC method was developed and applied to estimate the number of serious road injuries per vehicle type and to stratify this number in relevant subgroups. Information on serious road injuries was found in registries from both police and hospitals, which are both incomplete and contain misclassification. These variables were used as indicators of a latent variable of which it can be said that it contains the *true* scores. Posterior membership probabilities obtained from this latent class model were then used to create multiple imputations of these *true* scores. Simultaneously, multiple imputations were created for the missing values in the variable *region of accident* by using this variable as a perfectly measured indicator of the latent variable *region of accident* and supplementing it by specifying the variable *region of hospital* as an imperfectly measured indicator.

Multiple imputations were created for the variable *vehicle type* and for the variable *region of accident*. All variables are now fully imputed for every case in the dataset. Descriptive statistics of these variables, or estimates of relationships with other variables can now be investigated in a straightforward manner.

The extended MILC method was applied on datasets for the years 1994, 2009 and 2013. The quality of the data for these years was very different, which can be seen in the number of observations per registry per year and which is reflected in the entropy R^2 of the corresponding latent class model. In general the quality of the data was sufficient for applying the MILC method. In contrast, the results of the extended MILC method were compared to the results obtained when the hierarchical assignment procedure was applied (traditionally used to generate these statistics). A clear difference was that the extended MILC method generated wider 95% confidence interval widths. Based on the results obtained from the simulation study performed in Section 3.3.5, it can be concluded that these wider confidence interval widths were indeed necessary to obtain nominal coverage rates.

A number of issues are worth reflecting on a bit further. First of all, it is important to note that our results heavily depend on the model assumptions made. In particular, the assumption

is made that the classification errors are independent of covariates (also known as 'ICE' and 'homogeneous CE'). Furthermore, the assumption is made that the covariate variables are free of error. Violating this assumption does not necessarily have to be an issue if these errors are random. However, there is currently no literature on this topic, so more research in this specific area is needed in order to be able to adapt the model. A more crucial assumption is that the missingness is at random (MAR). Although from a theoretical perspective this assumption is likely to hold, it could however lead to substantial bias in cases where this assumption is violated.

A second issue is how the extended MILC method dealt with non-motorized vehicles. This was an ad hoc procedure to handle an issue that could not be handled by the latent class model. This ad-hoc procedure turned out to be useful. It can be investigated whether a comparable procedure could be applied to handle a moped/motorcycle issue in the 2013 dataset and whether there are other issues that can be solved like this.

This particular dataset contained a number of issues, of which a substantial part has been investigated by means of a simulation study. The results of this simulation study made clear that the extended MILC method was able to handle the missing values in the indicator variables and that the non-parametric bootstrap was required to obtain nominal coverage rates. It is however not investigated if and how large numbers of categories influence the results. Therefore, the number of imputations was increased and evaluated using methods to evaluate the number of imputations for missing values. A more thorough investigation could provide insight into whether these methods are suitable to evaluate the number imputations needed when the MILC method is applied, and how many imputations are needed to evaluate datasets with larger numbers of categories.

Furthermore, in the initial model proposed in Chapter 2, bootstrap samples were taken of the original data to incorporate parameter uncertainty in the estimate of the total variance. This appeared to be problematic for larger models with many interactions as those used in our application, because not all parameters can be estimated for every bootstrap sample. Alternatives to incorporate parameter uncertainty can be Bayesian MCMC or a parametric bootstrap. However, it should also be investigated whether such a step is still necessary for larger sample sizes as parameter uncertainty can become minimal in such cases. As the simulation study showed that it was necessary to incorporate parameter uncertainty when creating imputations for this specific case, a model with only main effects was used to enable estimation of all parameters.

At last, it is important to note that missing values in the combined dataset and classification errors in the observed data are not the only issues when estimating the total number of serious road injuries per vehicle type. There is also a number of serious road injuries that are neither observed by the hospital nor by the police. Weighting and capture/recapture methods are typically used to obtain an estimate of the total number of serious road injuries; approaches which can easily be combined with MILC by applying the methods on the imputations separately. A variance estimate would then include uncertainty about the total number of injuries which is typically estimated by making use of bootstrapping. This can also be applied separately to every imputation before pooling of the results is applied (Gerritse et al., 2016).

By creating multiple imputations using a latent class model, multiply imputed versions of variables that contained missing values and/or classification errors are created. These can be used to easily provide frequencies, to further divide these frequencies into relevant subgroups or to create statistical figures. This application showed that the initial MILC method can be extended to handle problems that are dataset-specific. Furthermore, this application highlighted various new problems that one may need to deal with when applying the MILC approach. In future research, these will be investigated more thoroughly to fully exploit the potential of the MILC method for dealing with classification error problems.

Chapter 4

Combining Multiple Imputation and Latent Markov modeling to obtain consistent estimates of *true* employment status

Abstract

Recently, a method was proposed that combines multiple imputation and latent class analysis (MILC) to correct for misclassification in combined datasets. A multiply imputed dataset is generated which can be used to estimate different statistics of interest in a straightforward manner, and which ensures that uncertainty due to misclassification is incorporated in the estimate of the total variance. In this paper, MILC is extended by using latent Markov modeling, so that it can handle longitudinal data and correspondingly create multiple imputations for multiple time-points. As recently many researchers have investigated the use of latent Markov modeling to estimate *true* employment rates using a combined dataset consisting of data originating from the 'Labor Force Survey' (LFS) and register data, this combined dataset is used for the set-up of the simulation study performed in this paper. Furthermore, the proposed method is applied to an Italian combined LFS-register dataset. We demonstrate how the MILC method can be extended to create imputations of *true* scores for multiple time-points and thereby show how the method can be adapted to practical situations.

This chapter was resubmitted after revisions

4.1 Introduction

Latent Class Analysis (LCA) is traditionally considered as a method to identify a categorical latent variable, whose categories are called latent classes, using categorical observed variables (McCutcheon, 1987, p.7), where the latent variable generally corresponds to a target phenomenon which could be measured in principle, but for practical reasons it is not. The idea underlining LCA can been extended to evaluate classification errors for categorical response variables, when different measures of the same phenomenon are available (Biemer, 2011, p.13). In this context, the latent variable is assumed to be the *true* score of the target phenomenon and the available sources are used as contaminated indicators of it. Generally, in such a situation, the number of classes of the latent variable are set to be equal to the number of categories in the indicator variables. The model output then provides information on how response patterns of the indicators are related to scores on the latent variable, that can be interpreted as the classification error of single categories within the indicators. It also provides information on the distribution of the latent variable given a response pattern on the indicator variables which can be used to estimate the value of the scores of the latent variable for each response pattern profile.

Moreover, because of the of the increased availability of different sources of information not directly collected for statistical purposes, LCA for classification error correction is increasingly applied within the field of official statistics. In this context, it has been applied on different research topics, such as neighborhood of residence (Oberski, 2015), home ownership (Chapter 2) and serious road injuries (Chapter 3). However, LCA for classification error correction emerges substantively within the field of employment research (Biemer, 2004; Kreuter et al., 2008; Magidson et al., 2009; Manzoni et al., 2010).

Labor Force Survey (LFS; Eurostat, 2012) is the survey used to estimate the employment rate, as well as its changes over time. In many European countries, administrative data are collected on a regular basis (this can vary from daily to yearly collection, depending on the country and the type of employment). A lot of research has been done on integrating surveys and administrative data to estimate not only the *true* employment rates, but also more detailed characteristics such as on *true* employment contract types (Pavlopoulos & Vermunt, 2015).

In this context, the use of latent Markov models (LMM) especially gains traction for estimating classification errors in panel data. In LMM, the longitudinal target variable is a latent process and the longitudinal response variables are contaminated measures of it. The LMM can be described as a two components model: the latent model describing the distribution of the

latent process, assumed to follow a Markov chain with a certain number of states, and the measurement model that describes the distribution of the response variables given the latent process. Covariates may be present in this model either in the measurement model or in the latent model. For each time occasion it is possible to have univariate longitudinal data, that is one response variable, or multivariate longitudinal data, that is more than one response variable. However, LMMs are more adaptable than LCA models because estimation of parameters does not necessary require multivariate longitudinal data for identifiability. Under the basic assumptions of LMM, i.e. the Markov property, conditional independence between classification errors, and time-homogeneous error probabilities, the model is identifiable with one response variable and a minimum of three panel waves. The identifiability of more complex models, for instance correlated error models, require the availability of multiple indicators (Bassi, 1997). Like LCA, the longitudinal model provides an estimate of the the classification error of single categories within the indicators over time together with the joint conditional probabilities of the latent status given the observed data, which can be used to predict the scores of the latent variable, as well as its changes over time.

Before these LMMs can be used to produce official statistics by National Statistical Institutes (NSI's), thorough investigation on limitations and sensitivity to the various assumptions that are made is essential. With this in mind, research has already been performed on the re-use of obtained parameter estimates (Pankowska et al., 2017b), the influence of linkage error (Pankowska et al., n.d.) and the influence of mixed mode survey designs (Pankowska et al., 2018). Furthermore, the LMM for estimating employment rates using combined LFS-administrative data is tailored to country specific issues. For example Pavlopoulos & Vermunt (2015) assume a serial correlation in the Dutch register data on employment, while Filipponi et al. (2019) do not make this assumption for their application to the Italian situation. Here, the LMM method is utilized to predict the employment status within the Italian employment register through the assignment of scores to the latent variable.

An interesting use of latent variable models in official statistics involves the assignment of scores to the latent variable. This assignment can for example be done by drawing from the posterior membership probabilities, which describe the probability of belonging to a specific latent class conditional on a profile of scores on indicators and covariates. Obviously, the uncertainty linked to this process has to be taken into account, especially when performing further statistical analyses. This can be done by making use of multiple imputations (Rubin, 1987). A combination of latent class modeling and multiple imputation has been proposed in Chapter 2 and has been denoted as the Multiple Imputation of Latent Classes (MILC) method.

More specifically, the MILC method uses a latent class model for variables (indicators) measuring the same latent construct (*true* variable) to estimate the posterior conditional probabilities of the latent status given the observed data. Next, the posterior membership probabilities obtained from the LCA model are used to create multiple imputations of the construct under investigation. Assigning values to the latent variable can be beneficial for a number of reasons. First, imputations can also be created for individuals that had missing values on either one of the indicator variables. Second, as imputations are created for the complete population, it becomes straightforward to produce consistent small-area estimates, to create cross-tables with different covariates or to create graphs. Third, because of the fact that multiple imputation is used, all results can be supplemented with appropriate variance estimates.

Therefore, the aim of this paper is to investigate how the MILC method can be adapted in such a way that scores are assigned to a *true* variable for multiple time-points. In Section 4.2, the latent Markov model developed by Filipponi et al. (2019), used as a starting point for this research, is described in more detail. Next, it is described how multiple imputations can be created using this model and the total procedure, from creating imputations to obtaining estimates, is described in detail. In Section 4.3, the performance of the imputation procedure proposed is investigated by means of a simulation study. At last, the imputation procedure is applied to three different regions in Italy using data from 2014 (Section 4.4). Section 4.5 concludes the work.

4.2 Methodology

The aim of this section is to describe in detail how multiple imputations of *true* employment scores can be made using LMM output. First, the LMM used for this investigation is introduced. Second, it is explained how multiple imputations of *true* employment scores can be created using this model. Third, the procedure that is followed to create the imputations is described step-by-step.

4.2.1 Latent Markov Model estimating *true* employment status in Italy

The LMM described in this section is developed by Filipponi et al. (2019). The available information is a person-linked combined dataset containing monthly employment status measured by administrative sources and by the LFS. The administrative data contains individual scores for the complete population for every month, while the LFS is administered

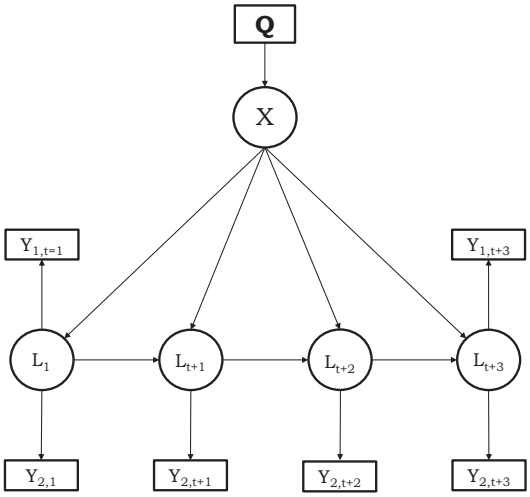

Figure 4.1: LMM used to estimate *true* employment rates per month in Italy, as developed by Filipponi et al. (2019)

twice a year per respondent, with three months in between. Despite that LFS data are only observed on a sample of the population, the classification errors of survey data are expected to be lower than those of administrative data. Because the aim is to predict the employment status within the Italian employment register, the choice between competing models has been based on the criteria of lower entropy, which should guarantee a lower classification uncertainty at the aggregate and at the individual level. The availability of indicators with low classification error helps the entropy reduction. The fitted LMM considers two indicators, both administrative and sample data, and 12 time-points, one for every month, spanning a time-frame of one year.

Figure 4.1 shows a graphical overview of the model. Here, let $Y_{1,t}$ denote the response variable originating from the LFS and $Y_{2,t}$ denote the response variable originating from the administrative source, where for both response variables $t = 1, ..., T$ and $T = 12$. The two vectors with elements $Y_{1,1}, ..., Y_{1,T}$ and $Y_{2,1}, ..., Y_{2,T}$ are respectively denoted by $\mathbf{Y_{(1)}}$ and $\mathbf{Y_{(2)}}$. The number of categories of the two response variables is equal, namely two, with (1) unemployed and (2) employed.

The vector $\mathbf{L} = (L_1, ..., L_T)$ represents the latent Markov variable measuring the *true* employment scores over time. Here, the number of latent states is equal to two, with (1) unemployed and (2) employed. A discrete (latent) random effect X is included in the model

to account for unobserved heterogeneity, so $\mathbf{L}$ follows a first order Markov chain conditional on X. In particular, the latent variable X identifies three subpopulations of individuals with different trajectories of $\mathbf{L}$ that can be described as never employed $(x = 1)$, always employed $(x = 2)$ and moving between employment and unemployment $(x = 3)$. Because of this mixture component, we can denote the model as a mixture latent Markov model.

The probability mass function of the latent variable X is affected by a number of time-invariant covariates, denoted by $\mathbf{Q} = (Q_1, Q_2, ..., Q_p)$. Multiple administrative sources were linked on person level to construct $Y_{2,t}$, and the covariate Q_1 identifies these different administrative sources, which are strongly related to the different types of employment contracts. The other covariates used in the model are: *retirement status*, *student*, *earnings*, *age* and *gender*. The use of the $\mathbf{Q}$ covariates will help in the identification of the different components of the latent variable X and therefore the different trajectories of $\mathbf{L}$. For more information on LMM utilized to predict the Italian employment status we refer to Filipponi et al. (2019).

For illustration purposes, we consider two covariates, Q_1 and Q_2, with m and n categories respectively. Therefore, corresponding model parameters are specified as:

$$\phi_{x|\mathbf{q}} = \mathrm{P}(X = x \mid Q_1 = q_1, Q_2 = q_2), \qquad \begin{aligned} x &= 1, 2, 3, \\ q_1 &= 1, \ldots, m, \\ q_2 &= 1, \ldots, n. \end{aligned}$$

The initial probabilities are specified as:

$$\pi_{l_1|x} = \mathrm{P}(L_1 = l_1 \mid X = x), \qquad \begin{aligned} l_1 &= 1, 2, \\ x &= 1, 2, 3. \end{aligned}$$

The transition probabilities are specified as:

$$\pi_{l_t|l_{t-1},x} = \mathrm{P}(L_t = l_t \mid L_{t-1} = l_{t-1}, X = x), \qquad \begin{aligned} l_t &= 1, 2, \\ x &= 1, 2, 3, \\ t &= 1, ..., T, \end{aligned}$$

and the conditional response probabilities are specified as:

$$\psi_{y_{j,t}|l_t} = \mathrm{P}(Y_{j,t} = y_t \mid L_t = l_t), \qquad \begin{aligned} y_t &= 1, 2, r \\ l_t &= 1, 2, \\ t &= 1, ..., T, \\ j &= 1, 2. \end{aligned}$$

For convenience, all expressions denoting realizations of random variables (e.g. x or l_t) are suppressed, unless for special situations. Note that only the latent variable X depends on Q_1 and Q_2 and that only the initial and latent transition probabilities depend on X. Overall, the distribution of the observed indicators, given the covariates is

$$\mathrm{P}(\mathbf{Y_{(1)}}\mathbf{Y_{(2)}} \mid \mathbf{Q}) = \sum_{x=1}^{3} \sum_{l_1=1}^{2} \sum_{l_2=1}^{2} \dots \sum_{l_T=1}^{2} \phi_{x|\mathbf{q}} \pi_{l_1|x} \prod_{t=2}^{T} \pi_{l_t|l_{t-1},x} \prod_{t=1}^{T} (\psi_{y_{1,t}|l_t})^{\delta_t} \psi_{y_{2,t}|l_t}, \qquad (4.1)$$

where δ_t indicates whether an observation for $Y_{1,t}$ is present at time-point t. When fitting the LMM to the person-linked combined dataset as described above, a number of assumptions are made. In defining the probability distribution in Equation 4.1, different assumptions are made. The first assumptions are made on the latent process. It is assumed that $\mathbf{L}$ is a homogeneous first order Markov Chain, that is a person's *true* employment at time-point t given its *true* employment at $t-1$, is independent of its *true* employment on $t-2$ and the latent transition probabilities do not change over time (Biemer, 2011, p.272). Therefore, the model retains the Markov assumption given the covariates and the latent variable controlling for unobserved heterogeneity. The assumptions made on the measurement model are the following. The classification errors of the indicators are locally independent and independent over time, given the latent variables (ICE, Singh & Rao, 1995), which can be considered as an extension of the local independence assumption made when applying LCA. Moreover it is assumed that the amount of classification error within the indicators does not change over time. At last, it is assumed that the missing values due to the panel construction are Missing Completely At Random (MCAR, Rubin, 1976) and missing values due to attrition are Missing At Random (MAR, Rubin, 1976) (Pavlopoulos & Vermunt, 2015).

4.2.2 Creating multiple imputations

The aim of fitting the model to the person linked data set is to impute the *true* employment status and this can be carried out by drawing from the posterior probabilities. We distinguish

between conditional imputation, where sequences of latent states are generated from the joint conditional probabilities given the observed data, $\mathrm{P}(L_t \mid L_{t-1}, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q})$, and marginal imputation, where the latent status is generated from the posterior probabilities, $\mathrm{P}(L_t \mid \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q})$, for $t = 1, .., 12$. The choice between marginal and conditional imputation depends on the aim of the researcher. If the entire sequence of the latent status is of interest, conditional imputation is more suitable, as an imputation of the latent status at time-point t is made conditional on the imputation made for time-point $t-1$. Alternatively, if the imputation of only one time-point is of interest, marginal imputation is more straightforward and could produce more efficient results if other parameters are of interest, such as latent transition probabilities.

Since the initial and latent transition probabilities of the latent process depend on the latent variable X, the marginal imputation of L_t and conditional imputation of L_t given L_{t-1} have to be carried out conditionally on X. Therefore, conditional imputation of L_t requires an imputation of X first, which is generated by sampling from the posterior membership probabilities:

$$\mathrm{P}(X = x \mid \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q}) = \frac{\mathrm{P}(\mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X = x, \mathbf{Q})\mathrm{P}(X = x \mid \mathbf{Q})}{\sum_{s=1}^{3} \mathrm{P}(\mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X = s, \mathbf{Q})\mathrm{P}(X = s \mid \mathbf{Q})}. \tag{4.2}$$

The conditional imputation of L_t conditional on the imputation of X (denoted by x^*) is then generated by sampling from the probabilities:

$$P(L_1 \mid X = x^*, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q}) = \frac{P(L_1, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mid X = x^*, \mathbf{Q})}{P(\mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X = x^*, \mathbf{Q})} \tag{4.3}$$

when $t = 1$ and

$$\mathrm{P}(L_t \mid L_{t-1}, X, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q}) = \frac{\mathrm{P}(L_t, L_{t-1} = l^*_{t-1}, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mid X = x^*, \mathbf{Q})}{\mathrm{P}(L_{t-1} = l^*_{t-1}, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X = x^*, \mathbf{Q})} \tag{4.4}$$

when $t > 1$ (which is also conditional on the imputation of L_{t-1} denoted by l^*_{t-1}) and where the distribution defined in Equation 4.2, Equation 4.3 and Equation 4.4 can be obtained by marginalizing

$$\mathrm{P}(\mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{L} \mid X) = \pi_{l_1|x} \prod_{t=2}^{T} \pi_{l_t|l_{t-1},x} \prod_{t=1}^{T} (\psi_{y_{1,t}|l_t})^{\delta_t} \psi_{y_{2,t}|l_t}.$$

Alternatively, the marginal imputation of L_t given X can be generated by sampling from the

posterior probabilities regardless of t, x^* or l^*_{t-1}:

$$\mathrm{P}(L_t \mid X, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}}, \mathbf{Q}) = \frac{\mathrm{P}(L_t, \mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X, \mathbf{Q})}{\mathrm{P}(\mathbf{Y_{(1)}}, \mathbf{Y_{(2)}} \mid X, \mathbf{Q})}.$$

It is important to underline that the evaluation of the quantities expressed in Equation 4.1 to Equation 4.4 involve sums over a large number of configurations. For example, to compute Equation 4.1, it is necessary to evaluate a sum over all possible 3×2^T configurations of the vectors $\mathbf{l}$ and X. An efficient way to compute the posterior membership probabilities is the forward recursion algorithm (Baum et al., 1970), which is implemented in Latent GOLD (Vermunt & Magidson, 2013b).

4.2.3 Multiple Imputation using the latent Markov Model

In the previous subsection, the LMM used for this research was described as well as the process to obtain its posterior membership probabilities of the latent status L_t for each time point. In this paragraph we describe the extension of the MILC procedure, developed in Chapter 2 in the context of LCA, to LMMs. The final aim of the procedure is to obtain the estimation of the average number of employers by domains and their variability. The MILC procedure is composed of five steps. In the *first* step, m non-parametric bootstrap samples are generated from the original dataset containing the indicators and covariates used to estimate the LMM. A bootstrap sample is obtained by sampling from the observed frequency distribution. In the *second* step, the LMM described in Section 4.2.1 is fitted on each of the m bootstrap samples. Then in the *third* step, one imputation for $\mathbf{L}$ is created using the m^{th} LMM obtained using the m^{th} bootstrap sample, resulting in m imputations of the sequence of $\mathbf{L}$. These imputations can be created using either the conditional imputation procedure or the marginal imputation procedure, as described in section 2.1. In the *fourth* step, estimates of interest can then be obtained from every imputation and in the *fifth* step, the estimates obtained for every imputation can be pooled using the pooling rules defined by Rubin (1987, p.76). For an example of how to apply the pooling rules in the MILC context to obtain pooled estimates of frequency tables, we refer to Chapter 2. Note that when the relationship between a covariate and the imputed *true* scores is of interest, the assumption is made that the covariate is free of error. It is important to note that drawing bootstrap samples in step one allows us to indirectly take into account parameter uncertainty through the imputations created in a later step. Therefore, not one LMM model based on one dataset is used, but m LMM models based on m bootstrap samples.

4.3 Simulation study

4.3.1 Set-up of the simulation study

The performance of the multiple imputation procedure using LMMs to obtain *true* employment scores is empirically evaluated by a simulation study. Since the LMM contains many parameters and there are many different ways in which the imputation procedure could be approached, the decisions made for this simulation study are exemplified in the consecutive subsections.

4.3.1.1 Population model

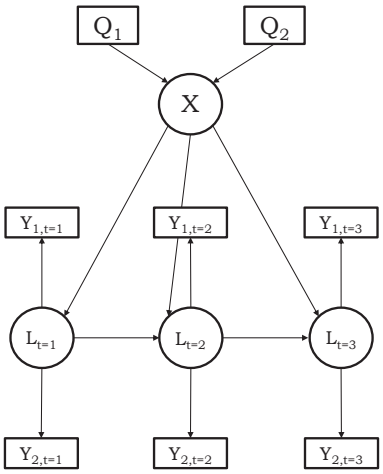

Figure 4.2: LMM used in the simulation study to evaluate the performance of the latent Markov multiple imputation procedure

The model used to draw samples has been specified to mimic as much as possible the combined Italian LFS-admin dataset (see Filipponi et al. (2019)). The specified model is described in Figure 4.2. In the simulation study, there are only three time-points observed for computation-time reasons.

When describing the model parameters, we start with the mixture variable X. As described, this variable is included into the model to take individual heterogeneity into account. In particular, we interpret the groups of people identified by X with people who are 'never employed' ($x = 1$), 'always employed' ($x = 2$) or 'move between being employed and unemployed' ($x = 3$). Two covariates are included in the model: Q_1 with the purpose to

mimic the influence of the *source* covariate (described in Section 4.2.1) and therefore has a strong relationship with X; Q_2 which has no relationship with the mixture at all. Table 4.1 shows the parameters ϕ_x and $\phi_{x|\mathbf{q}}$. The reason for including the covariate Q_2 is that it

Table 4.1: Parameters of latent variable X

ϕ_x									
	$x=1$			$x=2$			$x=3$		
	0.5800			0.3025			0.1175		
$\phi_{x\|\mathbf{q}}$									
	$q_2=1$	$q_2=2$	$q_2=3$	$q_2=1$	$q_2=2$	$q_2=3$	$q_2=1$	$q_2=2$	$q_2=3$
$q_1=1$	0.95	0.025	0.025	0.95	0.025	0.025	0.95	0.025	0.025
$q_1=2$	0.025	0.95	0.025	0.025	0.95	0.025	0.025	0.95	0.025
$q_1=3$	0.025	0.025	0.95	0.025	0.025	0.95	0.025	0.025	0.95

will be used later on for developing a MAR missingness mechanism. Furthermore, it would be interesting to compare the performance of the imputation procedure for estimating the relationship with a strongly related covariate in comparison to a weakly related covariate.

The set of three initial probabilities and transition probabilities of the Markov chain, for each mixture groups are reported in table 4.2 and table 4.3. Here it can be seen that the group

Table 4.2: Parameters of the latent Markov variable **L**: initial probabilities

$\pi_{l_1\|x}$		
	$l_1=1$	$l_1=2$
$x=1$	0.97	0.03
$x=2$	0.06	0.94
$x=3$	0.50	0.50

Table 4.3: Parameters of the latent Markov variable **L**: transition probabilities

$\pi_{l_t\|l_{t-1},x}$						
	$x=1$		$x=2$		$x=3$	
	$l_t=1$	$l_t=2$	$l_t=1$	$l_t=2$	$l_t=1$	$l_t=2$
$l_{t-1}=1$	0.97	0.03	0.94	0.06	0.70	0.30
$l_{t-1}=2$	0.03	0.97	0.06	0.94	0.30	0.70

'never employed' ($x=1$) has a strong relationship with $L_1=1$. In contrast, $x=2$ has a strong relationship with $L_1=2$, the 'employed' group in the Markov chain. As $x=3$

group moves from being employed and unemployed, their probability of being unemployed or employed on $t = 1$ is 0.5.

It can also be seen that for $x = 3$ the probability of changing from being unemployed to being employed or the other way around is higher compared to the other groups. At last, the error probabilities of the two indicator variables are specified in table 4.4: These

Table 4.4: Parameters of the measurement model

$\psi_{y_j \mid l}$				
	Condition 1		Condition 2	
	$l_t = 1$	$l_t = 2$	$l_t = 1$	$l_t = 2$
$y = 1$	0.95	0.05	0.80	0.20
$y = 2$	0.05	0.95	0.20	0.80

probabilities indicate that for both indicator variables, 95% is correctly classified, which can be considered realistic for the LFS indicator, but low for the administrative data. To investigate the performance of the procedure with data of a lower quality, indicators are also simulated with 80% correctly classified.

4.3.1.2 Parameters under evaluation

In the previous section, all specified model parameters were discussed. In theory, it is possible to evaluate the performance of the imputation procedures by investigating all these parameters. This would result in a large amount of information, which is, however, not all relevant. By choosing parameters useful for investigating the performance of the imputation method, we should think of the reason why the imputations are made in the first place. As discussed in the introduction, applications of the LMM are already done and provide us much information. What it does not give us, is information on the Markov variable in relationship to covariates of interest. Therefore, it makes sense to investigate $\pi_{\bar{l}|\mathbf{q}}$. Furthermore, $\pi_{\bar{l}}$ and ϕ_x are investigated as they both might be of interest to employment researchers. Here, ϕ_x can be found in Table 4.1, while $\pi_{\bar{l}}$ is obtained by the following marginalization:

$$\pi_{\bar{l}} = \frac{\sum_{q_1=1}^{m} \sum_{q_2=1}^{n} \sum_{x=1}^{3} \sum_{t=1}^{T} \phi_{x|\mathbf{q}} \pi_{l_1|x} \pi_{l_t|t-1,x} N_{q_1 q_2}}{\sum_{q_1=1}^{m} \sum_{q_2=1}^{n} N_{q_1 q_2} T}$$

and

$$\pi_{\bar{l}|\mathbf{q}} = \frac{\sum_{x=1}^{3} \sum_{t=1}^{T} \phi_{x|\mathbf{q}} \pi_{l_1|x} \pi_{l_t|t-1,x} N}{NT}.$$

The corresponding obtained population values can be found in Table 4.5.

Table 4.5: Parameters of latent variable $\pi_{\bar{l}}$

$\pi_{\bar{l}}$	$\bar{l}=1$	$\bar{l}=2$
	0.6388	0.3612

$\pi_{\bar{l}\|\mathbf{q}}$	$q_2=1$		$q_2=2$		$q_2=3$	
	$\bar{l}=1$	$\bar{l}=2$	$\bar{l}=1$	$\bar{l}=2$	$\bar{l}=1$	$\bar{l}=2$
$q_1=1$	0.9105	0.0895	0.9105	0.0895	0.9105	0.0895
$q_1=2$	0.1412	0.8588	0.1412	0.8588	0.1412	0.8588
$q_1=3$	0.5013	0.4987	0.5013	0.4987	0.5013	0.4987

4.3.1.3 Conditions under evaluation

As previously described, two alternative approaches for creating the imputations can be considered, conditional and marginal imputation are both evaluated in this simulation study. However, to evaluate different aspects of the proposed imputation procedures, multiple approaches for the entire imputation procedure are evaluated.

As a reference, the parameters described in Section 4.3.1.2 are obtained from the LMM output directly. To investigate the benefit of performing multiple imputations in comparison to a single imputation, parameters are also obtained after creating a single imputation. Furthermore, to investigate the extent of parameter uncertainty in the situation under evaluation, MILC is aplied both with and without bootstrap for parameter uncertainty. At last, to investigate the influence of the number of imputations, results are obtained using both five and ten imputations. Summarizing we compare 11 conditions:

- LMM: Obtaining results from the LMM output directly
- SI-C/M: Generate a single imputation (conditional or marginal) by estimating the LMM directly on the observed data.
- MI-5/10-C/M: Generate five or ten conditional or marginal imputations from the LMM estimated directly on the observed data.
- MI-B-5/10-C/M: Generate five or ten conditional or marginal imputations using five or ten LMMs estimated on five or ten bootstrap samples from the observed data.

4.3.1.4 Performance measures

The parameters under evaluation, as described in Section 4.3.1.2, are investigated using four performance measures. Note that we only provide the following equations for ϕ_x, while they can also be applied to the other parameters under evaluation, $\pi_{\bar{l}}$ and $\pi_{\bar{l}|\mathbf{q}}$. First, the bias of the parameters is investigated, which is equal to the difference between the average estimate over all replications and the value found in the theoretical population:

$$\text{bias}_{\phi_x} = \frac{\sum_{j=1}^{N_{\text{it}}} (\phi_x - \hat{\phi_x})}{N_{\text{it}}},$$

where N_{it}, stands for the number of simulation iterations performed in the simulation study, which is in this case always 500. Second, the coverage of the 95% confidence interval is under investigation and third, the ratio of the average standard error of the estimate over the standard deviation of the 500 replication estimates is estimated:

$$\frac{\left[\frac{\sum_{j=1}^{N_{\text{it}}}}{N_{\text{it}}} \text{SE}(\hat{\phi}_x)\right]}{\text{SD}(\hat{\phi}_x)}$$

where standard error (SE) is the square root of the estimate of the total variance obtained after applying the pooling rules by Rubin (1976) and

$$\text{SD}(\hat{\phi}_x) = \sqrt{\frac{\sum_{j=1}^{N_{\text{it}}} (\hat{\phi_x} - \bar{\phi_x})^2}{N_{\text{it}}}},$$

which is estimated to confirm that the standard errors of the estimates are properly estimated. At last, the root mean squared error is estimated:

$$\text{RMSE}_{\phi_x} = \sqrt{\frac{\sum_{j=1}^{N_{\text{it}}} (\hat{\phi_x} - \phi_x)^2}{N_{\text{it}}}}$$

4.3.2 Simulation results

In this paragraph the results of the simulation study are presented. In Figures 4.3 and 4.3: (i) the different rows of graphs represent three parameters investigated, (ii) the different rows within each graph represent different approaches to the longitudinal extension of the hidden Markov model, (iii) the four different combinations of colour and shape represent the four different simulation conditions.

In this section, only the simulation results graphically represented in Figures 4.3 and 4.4 are

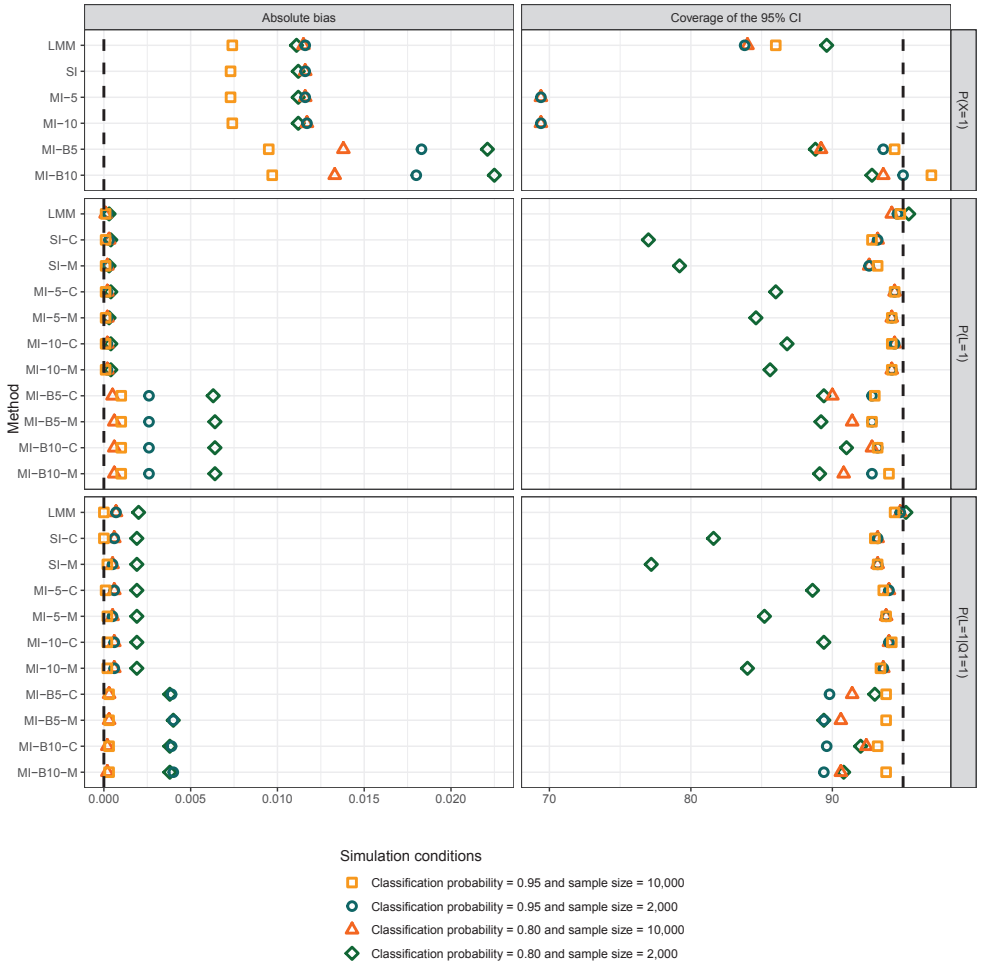

Figure 4.3: Plot of results in terms of bias and coverage of the 95% confidence interval in the columns. Note that we removed the following values from the results in terms of coverage: For SI: 42.6; 66.0; 66.0; 56.4 (in order of the conditions as listed in the legend). For MI-5: 45.6 (for $\psi_{1|1} = 0.80$ and ss=2000) and 61.2 45.0 (for $\psi_{1|1} = 0.95$ and ss=10.000). For MI-10: 45.6 (for $\psi_{1|1} = 0.80$ and ss=2000) and 60.4 (for $\psi_{1|1} = 0.95$ and ss=10.000)

discussed. Note that all parameters ϕ_x, $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$ behave in a similar way, so therefore only one parameter of each variable is graphically represented and discussed here. For the full set of simulation results, we refer to Appendix D.

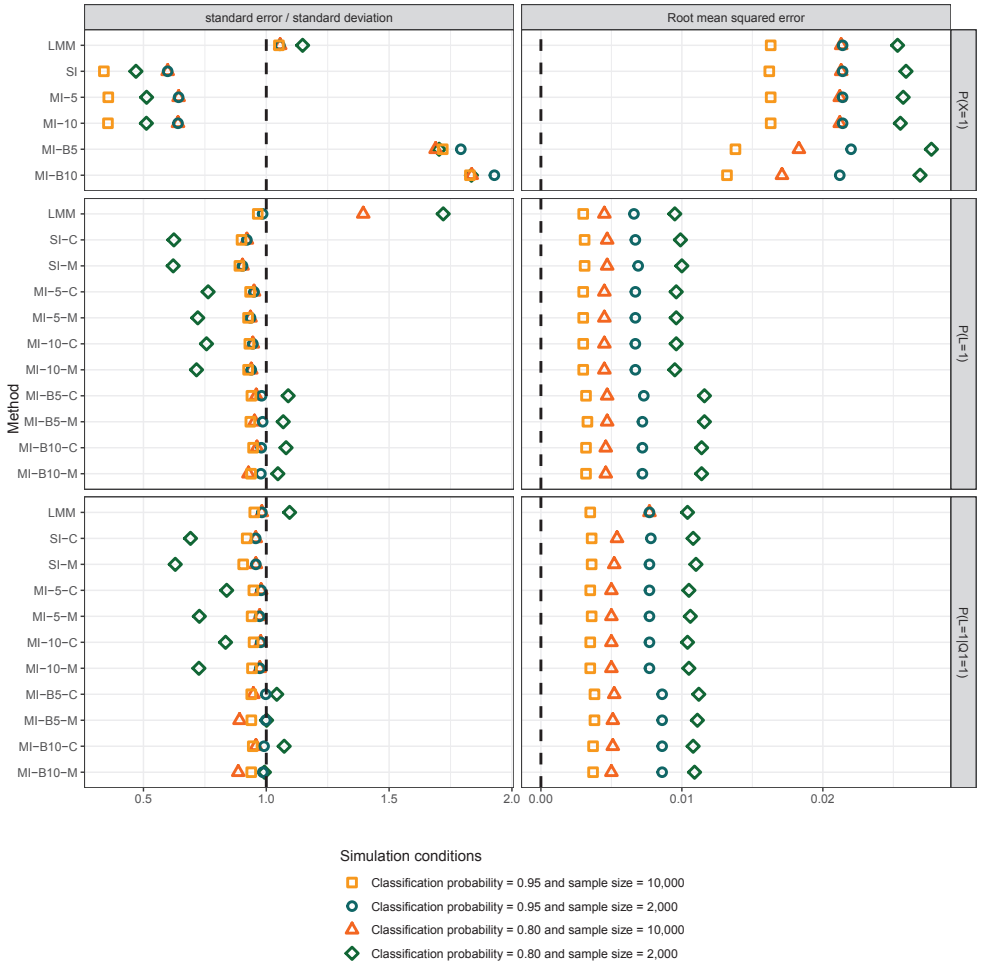

Figure 4.4: Plot of results in terms of SE/SD and RMSE

4.3.2.1 Latent Markov model results

When evaluating the results obtained directly from the hidden Markov model output, it can be seen that for $\pi_{\bar{l}=1}$ and $\pi_{\bar{l}=1|q_1=1}$ in Figure 4.3 and Figure 4.4 the model is able to produce estimates without bias and nominal coverage rates for all simulated conditions. The confidence interval width becomes wider as the simulation condition becomes more 'challenging' (smaller sample size and/or lower classification probability) and also the average standard error becomes larger in relation to the standard deviation over the estimates in such cases. In contrast, the hidden Markov model has more difficulties with estimating the

parameters of $\phi_{x=1}$. Here, a small amount of bias and undercoverage can be detected, and these amounts are related to the 'difficulty' of the simulation condition.

4.3.2.2 Single imputation versus multiple imputation

When comparing the results after a single imputation to the results after multiple imputation, it can be seen that in terms of bias of $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$, there are no problems for single nor multiple imputation. They also seem to perform well on other evaluation criteria for most simulation conditions. Coverage rates however are too low for both single and multiple imputation, but are worse for single imputation, and a similar pattern can be seen when evaluating the average standard error divided by the standard deviation over the estimates. For ϕ_x it can be said that both single and multiple imputation are not performing well. Here it is especially outstanding that the 'small sample size, large classification probability' and 'large sample size, low classification probability' conditions behave very similar, indicating that these characteristics behave in a sort of interchangeable way.

4.3.2.3 Bootstrapping for parameter uncertainty

In terms of bias, an increase can be seen when the bootstrap is applied in comparison to when it is not applied. However, it can also be seen that for the most difficult simulation condition (the condition with undercoverage for $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$), the results improve when the bootstrap is applied. This also results in a wider confidence interval and a SE/SD that is a bit larger than the nominal value of one compared to smaller (which was the case when no bootstrap was applied). The RMSE also indicates that the bootstrap results in a loss of efficiency. For ϕ_x, the bias and confidence interval width increase even more in comparison to $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$, but this also results in coverage rates changing from unacceptable to almost nominal.

At last, results obtained after creating different numbers of imputations was also investigated. However, the differences between creating 5 or 10 imputations are not noteworthy.

4.3.2.4 Conditional versus marginal imputation

Almost no differences can be seen between the two different imputation procedures. The only notable difference is found in the most difficult condition, which had undercoverage for both $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$. The results obtained after conditional imputation lead to a coverage rate closer to the nominal 95% level compared to the results obtained after marginal imputation.

4.3.3 Missing values

The main purpose of this paper is to investigate whether the longitudinal extension of the MILC method can be used to estimate employment rates using a combined dataset containing information from the Labor Force Survey (LFS) and administrative data. As the LFS only contains a sample of the population (and this subset also contains missing values), it makes sense to investigate if the missing values (both by non-response and by design) influence the quality of the estimates obtained when the longitudinal extension of the MILC method is applied.

4.3.3.1 Simulation conditions

As the simulation condition with classification probabilities of 0.95 and sample size of $10,000$ is closest to the situation where this model is applied in practice, only this condition will be used to further investigate the influence of missing values. In practice, only a very small subset of the population has observations on the LFS indicator variable measuring employment (see section 4 for some exemplary numbers). Although we would ideally mimic this situation, this would not be feasible for a simulation study for computational reasons. We decided to set the percentage of missing cases for the indicator representing the LFS to 50%. In that way, a substantive amount of cases is dropped, but we are still left with a reasonable amount of profiles. Both a Missing Completely At Random (MCAR) mechanism as a Missing At Random

Table 4.6: Illustration of the simulated data structure containing missing values

case ID	$Y_{1,t=1}$	$Y_{1,t=2}$	$Y_{1,t=3}$	$Y_{2,t=1}$	$Y_{2,t=2}$	$Y_{2,t=3}$	Q_1	Q_2
1	1	1	1	1	1	NA	1	1
2	1	1	1	1	1	NA	1	1
3	1	1	1	NA	NA	NA	1	1
4	1	1	1	NA	NA	NA	1	1

(MAR) mechanism are investigated. With the MCAR mechanism, the probability of being missing is equal for all respondents, so 0.50. With the MAR mechanism, the probability of being missing is not equal for all respondents. Instead, the probability of being missing is related to a respondent's score on another variable. In this case, it is related to the score of the

respondent on covariate Q_2. If

$$Q_2 = 1, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.25;$$
$$Q_2 = 2, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.50;$$
$$Q_2 = 3, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.75,$$

where the total number of missing cases also depends on the frequency distribution of Q_2. Since we are trying to mimic the structure of the longitudinal combined LFS-admin dataset, we use the missingness mechanism to generate 50% missing values on all time-points of Y_2. Furthermore, since the LFS is only observed for two time-points, the third timepoint of Y_2 is made missing for all observations. See Table 4.6 for an illustration of this data structure.

4.3.3.2 Simulation results

In Figures 4.5 and 4.6, an overview of the results obtained after applying the longitudinal extension of MILC on a dataset with MCAR or MAR missingness are compared to results obtained after applying the longitudinal extension of MILC on a fully oberved dataset. In Figure 4.5 and Figure 4.6, the rows of graphs represent three parameters investigated, while the rows within each graph represent different approaches to the longitudinal extension of the LMM. The three combinations of color and shape represent the MCAR, MAR and fully observed simulation conditions.

As in the simulation study conducted in Sections 4.3.1 and 4.3.2, it can be seen that the results for ϕ_x are more problematic compared to the other results. In general, the results for MCAR or MAR are very comparable to those obtained when a fully observed dataset is used. Only in terms of bias and in terms of RMSE, some differences can be found. Here, it can be seen that the bias and RMSE increase slightly when we shift from fully observed to MCAR, and increase a bit more when shifting from MCAR to MAR. In terms of $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$, the results with MAR and MCAR are even more similar to the fully observed results compared to those obtained for ϕ_x.

4.4 Application

The longitudinal extension of the MILC method is applied separately to data from 2014 of three different regions in Italy: Veneto, Umbria and Basilicata. As can be seen in Figure 4.7, the regions are spread out over the country, from north to south. Also, the regions differ

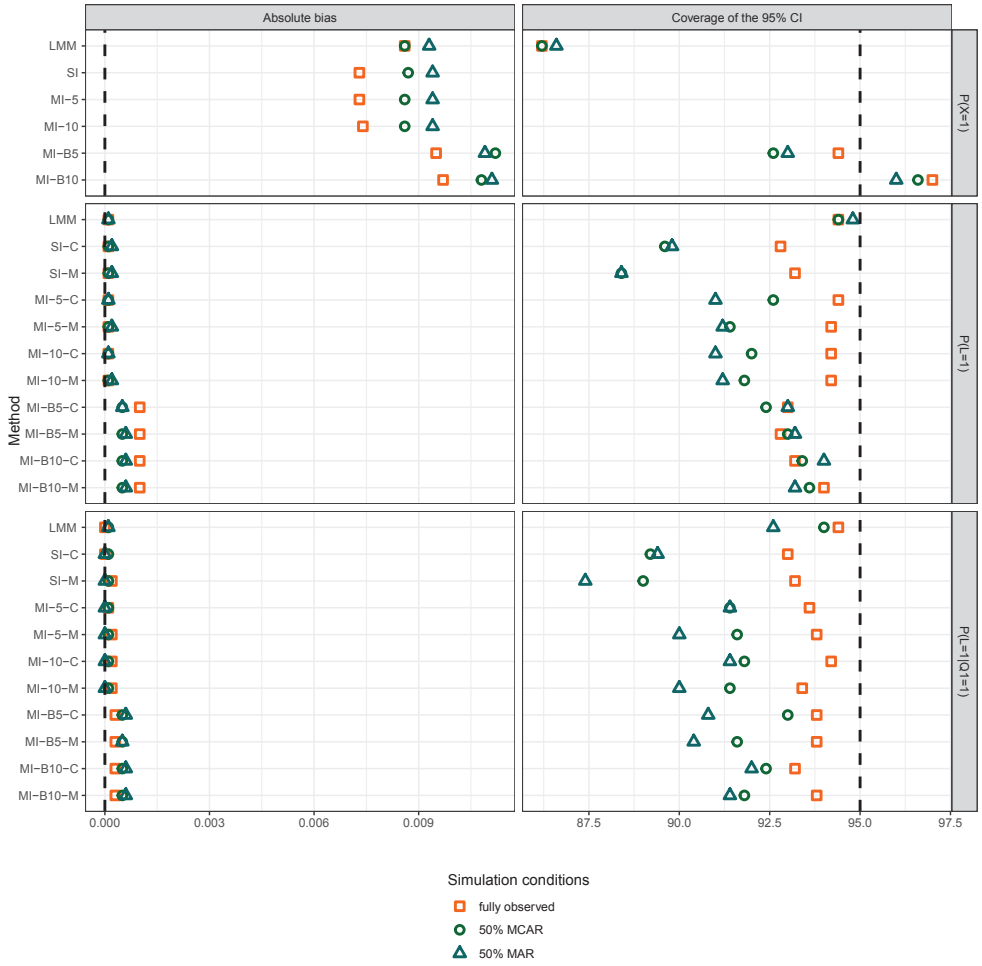

Figure 4.5: Plot of results in terms of bias and coverage of the 95% confidence interval in the columns. Note that we removed the following values from the results in terms of coverage: For SI: 42.6; 43.4; 43.2. For MI-5: 45.6; 45.8; 45.8. For MI-10: 45.0; 45.2; 45.8 (in order of the conditions as listed in the legend).

substantively in the number of inhabitants in the workforce and number of LFS respondents: Veneto has $4,821,983$ inhabitants in the workforce and $17,246$ LFS respondents; Umbria has $899,366$ inhabitants in the workforce and $8,477$ LFS respondents; Basilicata has $579,860$ inhabitants in the workforce and $10,202$ LFS respondents.

On these datasets, the longitudinal extension of the MILC method is applied using the LMM described in Figure 4.1 and Section 4.2.1. As described in Section 4.2.1, a covariate (Q_1) is included that specifies from which administrative *source* the $Y_{2,t}$ indicator score originates.

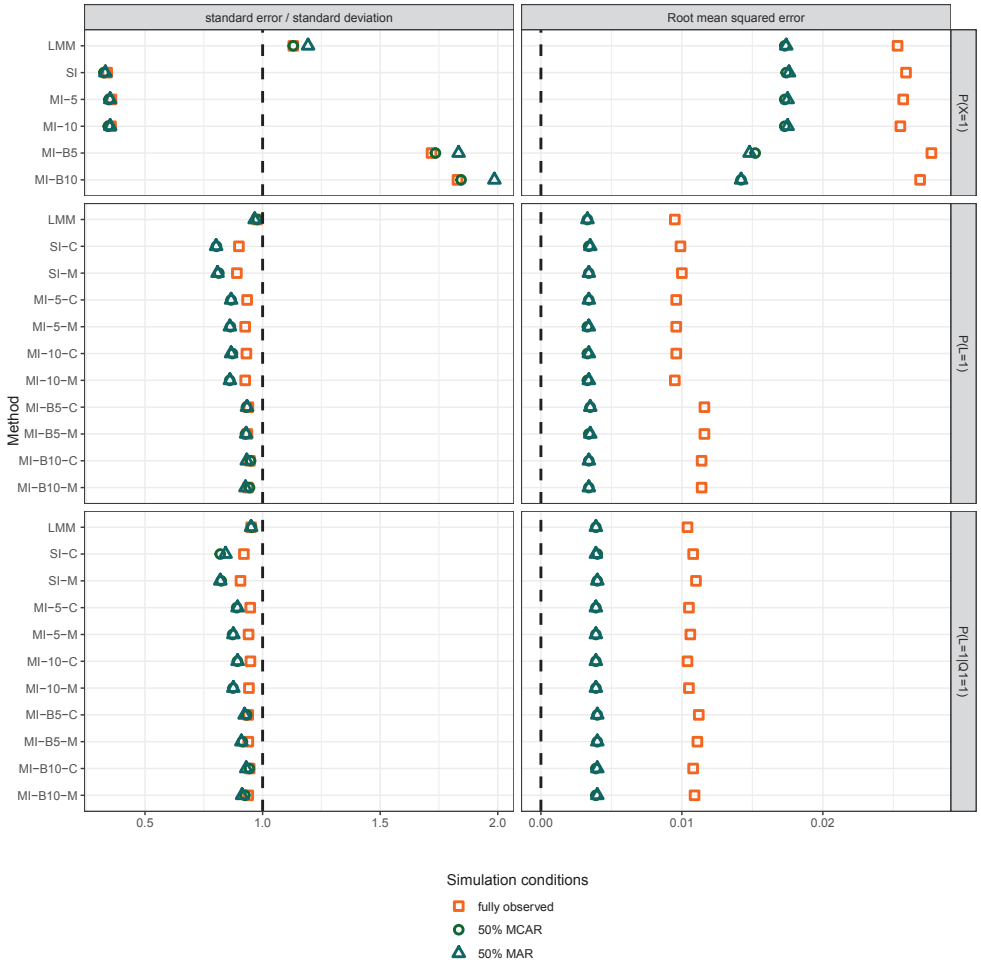

Figure 4.6: Plot of results in terms of SE/SD and RMSE.

Q_1 has four categories: No source (1); Employees (2); Self-employers, with time information (4); Self-employers, no time information (4). One other covariate is included in the model: *gender* (Q_2) with the categories Male (1) and Female (2).

Based on the results from the simulation study conducted in Sections 4.3.2 and 4.3.3, the longitudinal extension of the MILC method is applied using five bootstrap samples. As we concluded from the simulation study that marginal and conditional imputation produce similar results, only the results for conditional imputation are shown.

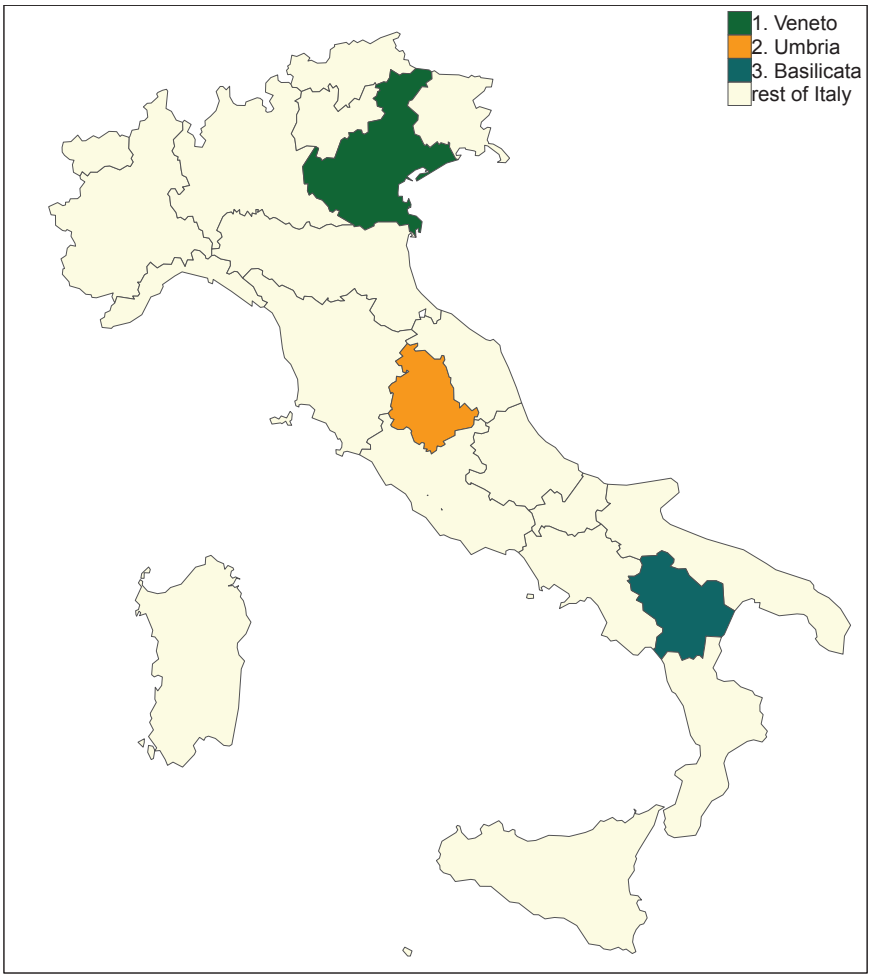

Figure 4.7: Map of Italy with the three regions highlighted on which the longitudinal extension of the MILC method is applied

In Table 4.7, the results in terms of proportions and corresponding standard errors for the different regions are found. By using proportions, it is possible to directly compare the employment rates over the different regions. Here, it can be seen directly that the employment rate ($\pi_{\tilde{l}=2}$) decreases as we shift to a more southern region. Similarly, we see the proportion of the mixture group representing the 'employed trajectory' ($\phi_{x=1}$) also decreases as we shift to a more southern region.

When we investigate the proportion of being employed conditional on the administrative source where a persons information was obtained, it can be seen directly that source one

Table 4.7: Results in terms of proportions obtained after applying the longitudinal extension of the MILC method to data from three different regions in Italy. The columns represent the estimates and standard errors of the different regions, and the rows represent the different parameters, which are the same as investigated in the simulation studies

	Veneto		Umbria		Basilicata	
	estimate	S.E.	estimate	S.E.	estimate	S.E.
$\phi_{x=1}$	0.37423	0.00009	0.33869	0.00036	0.26793	0.00028
$\phi_{x=2}$	0.53421	0.00007	0.57020	0.00017	0.61503	0.00022
$\phi_{x=3}$	0.09156	0.00008	0.09111	0.00039	0.11704	0.00027
$\pi_{\bar{l}=1}$	0.57994	0.00007	0.61601	0.00015	0.67801	0.00018
$\pi_{\bar{l}=2}$	0.42006	0.00007	0.38399	0.00015	0.32199	0.00018
$\pi_{\bar{l}=1\|q_1=1}$	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
$\pi_{\bar{l}=2\|q_1=1}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\pi_{\bar{l}=1\|q_1=2}$	0.23606	0.00008	0.24065	0.00021	0.42219	0.00033
$\pi_{\bar{l}=2\|q_1=2}$	0.76393	0.00008	0.75934	0.00021	0.57781	0.00033
$\pi_{\bar{l}=1\|q_1=3}$	0.10210	0.00003	0.11137	0.00012	0.17714	0.00017
$\pi_{\bar{l}=2\|q_1=3}$	0.89790	0.00003	0.88863	0.00012	0.82286	0.00017
$\pi_{\bar{l}=1\|q_1=4}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\pi_{\bar{l}=2\|q_1=4}$	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
$\pi_{\bar{l}=1\|q_2=1}$	0.50314	0.00003	0.55092	0.00008	0.59548	0.00008
$\pi_{\bar{l}=2\|q_2=1}$	0.49686	0.00003	0.44908	0.00008	0.40452	0.00008
$\pi_{\bar{l}=1\|q_2=2}$	0.65305	0.00003	0.67603	0.00006	0.75725	0.00007
$\pi_{\bar{l}=2\|q_2=2}$	0.34695	0.00003	0.32397	0.00006	0.24275	0.00007

does not contain any employed persons ($\pi_{\bar{l}=2|q_1=1}$), while source four does not contain any unemployed persons ($\pi_{\bar{l}=1|q_1=4}$). Furthermore, it can be seen that the proportion employed is particularly higher for the Basilicata region compared to the other regions in source two ($\pi_{\bar{l}=1|q_1=2}$).

When investigating the proportion of being employed conditional on *gender*, differences between north and south are also visible. For example, the proportion of being unemployed conditional on being male ($\pi_{\bar{l}=1|q_2=1}$) shifts from approximately 0.50 in Veneto to approximately 0.60 in Basilicata, while the proportion of being unemployed conditional on being female ($\pi_{\bar{l}=1|q_2=2}$) also increases if we shift from Veneto to Basilicata, from approximately 0.65 to approximately 0.75. So although the probability of being employed is larger for males compared to females, this relationship does not change over the different

regions.

4.5 Discussion

In this article, MILC was adapted in such a way that scores can be assigned to a *true* variable for multiple time-points. In the recent literature, Latent Markov Models have been increasingly used in the field of research on employment. In this context, LMMs are applied on unit-linked combined datasets where information comes from different data sources, potentially affected by classification errors. One possible use of LMMs is to create imputations of the *true* scores of the latent categorical variable for multiple time-points. This paper presented a new method to impute the *true* scores of the latent categorical variable, by correcting for classification error in all data sources and measuring and taking into account the uncertainty of the imputations. This method is an extension of the Multiple Imputation of Latent Classes (MILC) method that has been proposed in Chapter 2 in the context of LC analysis. This paper has shown that imputations for different time-points can be generated in multiple ways and thereby illustrates the flexibility of the MILC method. In particular, a simulation study highlighted the usability of the MILC method in different conditions.

A simplification of the LMM developed by Filipponi et al. (2019) was used in a simulation study, where results of multiple alternative strategies that could haven been chosen were evaluated. In the first simulation study, these strategies were compared using data of different sample sizes and of different quality. The main conclusion of this simulation study was that the results related to the Markov chain measuring employment status were of a different quality compared to the results related to the mixture measuring different trajectories of employment status over time. From this, we can conclude that if a researcher is interested in evaluating the mixture, it is necessary that the bootstrap is applied to incorporate parameter uncertainty into the estimate of the total variance. When the data is of sufficient quality (which was apparently already the case with 20% classification error), parameter uncertainty for the Markov model is at such a low rate that it can be ignored. Note that when multiple imputation is applied in this case without incorporating the bootstrap for parameter uncertainty, reliable results related to the mixture cannot be obtained.

The small differences between the five or ten imputations indicated that a low number such as five was probably enough. Furthermore, two different imputation procedures were evaluated (conditional and marginal) and the results showed minor advantages for conditional imputation. Marginal imputation was however much more straightforward to

apply since an imputation for every time-point could be created unconditional from the imputation of the other time-points, This is something to be taken into consideration when creating imputations of the LMM.

Furthermore, MCAR and MAR missingness mechanisms were investigated and did not show substantive reductions in the quality of the output, so the longitidinal extension of the MILC method should be able to handle the missingness structure that is present in the combined LFS admin data, since a combination of MAR and MCAR is assumed here.

For illustration purposes, the longitudinal extension of the MILC method was applied to three different regions of Italy. Most of the assumptions made when applying multiple imputation of the LMM were related to the LMM itself. This model has been thorougly investigated by many researchers, both inside and outside the field of official statistics and using data from multiple countries. However, these conclusions might not hold for when substantive changes within the investigated labor markets happen. For example, the introduction of a basic income could influence the sizes of the mixture groups and possibly also the number and type of mixture groups.

An assumption that has always been made when applying multiple imputation of latent classes, is the assumption that the covariates are free of classification error. Although this assumption is probably never met in practice, it is not a problem when a simple latent class model is used, as long as the classification error is random. However, more investigation should be done to see if this holds when the LMM is used, especially in the way the model is currently specified. In this model, the mixture groups are determined by the covariates, and classification error in a covariate might result in assignment to the wrong mixture group. In addition, throughout this paper the assumption of local independence is made and a model is used that implies that classification error in both the survey and administrative data is random. However, in real situations it can occur that the classification error is autocorrelated. In surveys, this is primarily due to individual characteristics and personal responding style, while in administrative data it might result from the fact that once an error is made it is likely to be copied onto the following time-point. Performance of combining LMM and multiple imputation when this assumption is violated has not been investigated in this paper.

It should also be noted that the model used for the simulation study, was in some ways a simplification of the model developed by Filipponi et al. (2019). The first simplification related to the number of time-points. The main reason for reducing this number, was that a large

number of time-points results in a large number of possible profiles, and these can result in parameters not being estimated when applying the bootstrap for parameter uncertainty. It would be interesting to investigate whether LMM with larger number of time-points can be investigated using alternative ways for estimating parameter uncertainty, such as a parametric bootstrap or the use of a Gibbs sampler. A second simplification relates to a number of edit restrictions that were specified in the model developed by Filipponi et al. (2019). Although previous research has shown that multiple imputation of latent classes is able to incorporate edit restrictions (Chapter 2), we decided to leave them out of this research because the number of different research settings was already extensive and the edit restrictions in this setting were very specific for the Italian situation.

In summary, the longitudinal extension of MILC presented in this article allows for imputation of *true* scores on multiple time-points for longitudinal data structures together with their variability.

PART III

Extending the MILC method using external covariates

Chapter 5

Updating latent class imputations with external auxiliary variables

Abstract

Latent class models are often used to assign values to categorical variables that cannot be measured directly. This *imputed* latent variable is then used in further analyses with auxiliary variables. The relationship between the imputed latent variable and auxiliary variables can only be correctly estimated if these auxiliary variables are included in the latent class model. Otherwise, point estimates will be biased. We develop a method that correctly estimates the relationship between an imputed latent variable and external auxiliary variables, by updating the latent variable imputations to be conditional on the external auxiliary variables using a combination of Multiple Imputation of Latent Classes (MILC) and the so-called three-step approach. In contrast with existing *one-step* and *three-step* approaches, our method allows the resulting imputations to be analyzed using the familiar methods favored by substantive researchers.

This chapter was published as: L. Boeschoten, D.L. Oberski, T. De Waal, J.K. Vermunt (2018) Updating latent class imputations with external auxiliary variables *Structural Equation Modeling: A Multidisciplinary Journal* 25 (5), 750-761

5.1 Introduction

In many different disciplines, multiple observed variables are used as indicators of one latent categorical variable that cannot be measured directly. For example in sociology, multiple indicators are used to distinguish latent classes of sexual morality and pro-life values (McCutcheon, 1987). In official statistics (the field of research concerned with the publishing of statistics for government or other official agencies), indicators from multiple sources are used to estimate the number of temporary and permanent employment contracts in the Netherlands (Pavlopoulos & Vermunt, 2015). In these settings, the latent variable of interest is estimated by including observed variables as indicators in a latent class (LC) model. This LC model is then used to assign values to the latent variable itself. This *imputed* latent variable (also known as a *plausible value* (Mislevy, 1991; Mislevy et al., 1992)) is often used in further analyses with auxiliary variables. For example, to relate different levels of sexual morality and pro-life values to attitudes towards abortion (McCutcheon, 1987) or to relate type of employment contract to level of education (Pavlopoulos & Vermunt, 2015).

The relationship between the imputed latent variable and auxiliary variables can only be correctly estimated if the auxiliary variables of interest are included in the LC model. Otherwise, point estimates will be biased (Wu, 2005; Monseur & Adams, 2009). This bias is due to the estimates being conditional on the imputed latent variable, and not on the latent variable itself (Bolck et al., 2004). Therefore, all auxiliary variables potentially of interest should be included in the LC model. However, this may not be possible or desired. For example in cases where an auxiliary variable is considered a distal outcome of the latent variable (Bakk, 2015, p.2). Another example is when the constructors of the measurement model do not want to share the indicator variables with the analysts due to privacy concerns. A third example is when the auxiliary variables are unavailable when constructing the measurement model due to a longitudinal or composite nature of the dataset.

Bias in the point estimates, caused by the absence of the auxiliary variables in the LC model, can be seen as a form of misclassification in the imputed latent variable. Therefore, methods that correct for misclassification should be considered, and we distinguish between different groups of methods. The first group of methods focuses on correcting the imputations of the latent variable and include Multiple Imputation for Measurement Error (MIME; Cole et al, 2006), Regression Calibration (RC; Spiegelman et al., 1997) and the complete re-estimation of the LC model (Schofield et al., 2014). For the latter, Multiple Imputation of Latent Classes can be used (MILC, Chapter 2). The advantage of these methods is that after correction, an

adapted dataset is produced that can be used to perform any type of analysis. The main drawback is that every time that new external auxiliary variables are acquired, complete re-estimation of the LC model is required. The second group of methods correct the estimate describing the relationship that is prone to bias. This group includes methods as simulation extrapolation (SIMEX; Cook & Stefanski, 1994) and the latent class three-step approach (Bolck et al., 2004). Their main advantage is that uncertainty due to misclassification is correctly incorporated into the estimates after new external variables are acquired. However, an important disadvantage is their inflexibility; a separate procedure needs to be followed for every analysis and a likelihood needs to be available to obtain the estimates of interest. Such complications prevent these important corrections from gaining traction among substantive researchers.

We develop a general approach by combining a method based on model correction (implementation of the LC model using the MILC method) with a method that is based on correction for bias (the three-step approach). This new approach preserves the advantages of both methods while discarding their disadvantages due to its generic nature. More specifically, this combined method (from now on denoted as the three-step MILC method) uses an LC model to create multiple imputations of the latent variable, which includes both parameter uncertainty and latent variable uncertainty into the estimate of the variance. Next, information from the LC model is used to estimate the amount of misclassification in the imputed latent variable. The estimate of this misclassification is then used to correct the relationship between the imputed latent variable and external auxiliary variables. Finally, the latent variable imputations are updated to be conditional on the external variables.

In the Section 5.2, issues currently faced by researchers are discussed in more detail, for which we present the three-step MILC method as a solution in Section 5.3. In Section 5.4, a simulation study is conducted to investigate the performance of the three-step-MILC method. In Section 5.5, the three-step-MILC method is applied on two empirical datasets, followed by a discussion Section 5.6.

5.2 Background

Researchers frequently summarize multiple observed variables $(Y_1, ..., Y_L)$ into one latent variable (X). A model $\mathrm{P}(\mathbf{Y}|X)$ is constructed to estimate the values of X. This model is used to assign estimated values to X, resulting in an imputed version of the latent variable, W. Different rules can be used to assign values to W using $\mathrm{P}(X|\mathbf{Y})$, such as modal (McLachlan,

1992), proportional (Dias & Vermunt, 2008) or random assignment. With the latter, individuals are assigned to classes by sampling from the posterior $P(X|\mathbf{Y})$, so $W \sim P(X|\mathbf{Y})$ (Bakk, 2015, p.11). Regardless of the method used for assigning values to W, W is never a perfect representation of X; some misclassification is always introduced (Bakk, 2015, p.12).

The imputed variable W is created so it can be used in further analyses with auxiliary variables ($\mathbf{Q}$). As addressed by Lanza et al. (2013) and implied by Blackwell et al. (2015), $P(X|\mathbf{Q})$ can only be correctly estimated using $P(W|\mathbf{Q})$ if $\mathbf{Q}$ is included in the model used to assign values to W. In other words, when the covariate-adjusted posterior $P(X|\mathbf{Y}, \mathbf{Q})$ is used to determine W. Otherwise, biased estimates for $P(X|\mathbf{Q})$ are obtained (Bartlett et al., 2015; Bolck et al., 2004; Schofield, 2014; Tanner & Wong, 1987), unless the measurement is perfect, such that $P(W|\mathbf{X}) = 1$ for exactly one value of X for each value of W (see also Marsman et al. (2016) for the same result in IRT).

Although this problem does not arise if $\mathbf{Q}$ is included in the LC model used when estimating X, we consider situations here where this is neither possible nor desired. For example, $\mathbf{Q}$ may not have been collected yet when $P(\mathbf{Y}|X)$ was estimated, or researchers may be resistant to include $\mathbf{Q}$ in the initial measurement model. As a result, $P(X|\mathbf{Y}, \mathbf{Q})$ is not available, only $P(X|\mathbf{Y})$ is. It is, however, possible to obtain information about the misclassification in W from the LC model, $P(W|X)$, which is estimated as a byproduct of the parameters in $P(\mathbf{Y}|X)$ and $P(X)$ and the chosen assignment rule (Bakk et al., 2013). These two pieces of information, $P(W|X)$ and $P(W|\mathbf{Q})$, can be combined to obtain an estimate for $P(X|\mathbf{Q})$ using maximum likelihood (Vermunt, 2010) or weighting (Bolck et al., 2004), which are both approaches of latent class three-step modelling. By specifying the log-linear model in its most general form, the newly imputed version of W can be used to estimate any type of relationship with $\mathbf{Q}$. Consequently, researchers do not have to think in advance about the kind of relationship to investigate at a later stage.

However, when a single imputation of W is created using this approach, uncertainty about X is not included in the estimate of the variance. Therefore, multiple imputations of W should be created so that the differences between the imputations reflect this uncertainty (Rubin, 1987, p.76). This approach is a combination of the MILC method used for model correction and the three-step approach used to correct for bias.

5.3 Methodology

In this section, we present a solution for the problem that biased estimates are obtained when an imputed latent variable is related to external auxiliary covariates. The methodology is discussed step by step, starting with the methodology of the MILC method (Chapter 2) followed by its three-step (Vermunt, 2010) extension.

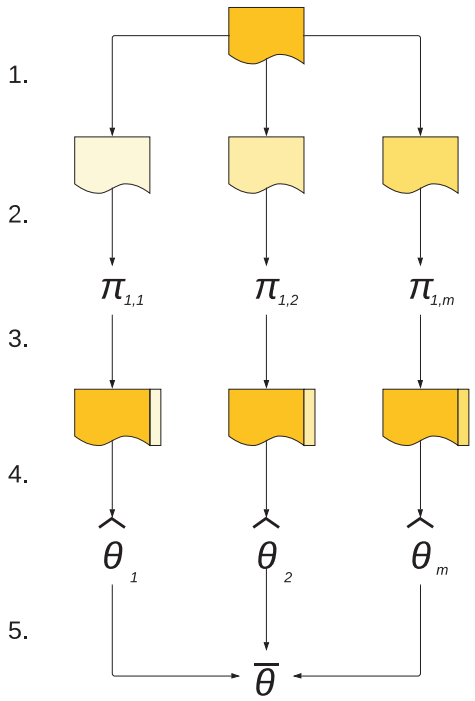

Figure 5.1: Graphical overview of the MILC method. The method starts with a dataset containing indicators and available covariate variables. At step 1, m bootstrap samples are drawn from the original dataset. At step 2, an LC model is built for each bootstrap sample (denoted by π). At step 3, m imputations for the latent variable are created. Estimates of interest are obtained from the m imputations, represented by $\hat{\theta}$ (step 4). Pooling these estimates to obtain $\bar{\theta}$ is the fifth step.

5.3.1 MILC

In Figure 5.1, a graphical overview of the MILC method is shown. The starting point of the method is a dataset comprising L indicator variables. In the *first* step, m bootstrap samples are drawn by sampling with replacement from the observed probability distribution of the original data, where m is equal to the number of multiple imputations created in a later stage. By using multiple imputations, we are able to include uncertainty due to measurement

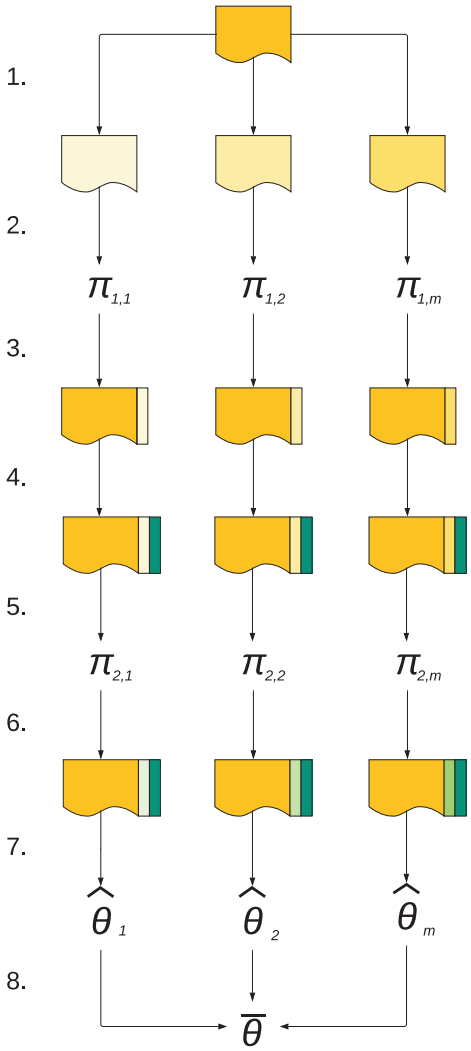

Figure 5.2: Graphical overview of the Three-step MILC method (using ML and BCH). All methods start with a dataset containing indicators and available covariate variables. Step 1, 2 and 3 are equal to the steps applied using the regular MILC method. With three-step MILC, step 4 is obtaining the classification table. Step 5 is to apply the ML or BCH correction procedure. Posterior membership probabilities are used to update the imputations for the latent variable (step 7). From the imputations, estimates can then be obtained and pooled (step 8)

error in the indicators when estimating the variance. By using bootstrap samples, parameter uncertainty is also included. This is especially recommended when datasets with smaller sample sizes are used as parameter uncertainty can be substantial in such cases (Wisz et al., 2008).

In the *second* step, an LC model is estimated for every bootstrap sample using the L indicator variables $(Y_1, ..., Y_L)$ of latent variable X, which has C categories denoted by $x = 1, ..., C$. C is equal over the bootstrap samples. If MILC is used to correct for measurement error in combined datasets, C is equal to the number of categories in the indicators. Available auxiliary variables can also be incorporated in the LC model as covariates and are denoted by $\mathbf{Z}$. The LC model for the probability of response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z})$ is then defined as:

$$\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = \sum_{x=1}^{C} \mathrm{P}(X = x \mid \mathbf{Z} = \mathbf{z}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x).$$

In some applications, one may wish to account for combinations of scores between the covariate variables and the latent variable that are not possible in practice. An example of such an impossible combination of scores is a Dutch person having *marital status* 'married' and *age* 'below 16 years', as this is prohibited by law. Edit restrictions are used to account for such impossible combinations of scores (De Waal et al., 2012), and can be specified in the LC model:

$$\mathrm{P}(X = \text{'married'} \mid Z = \text{'age below 16'}) = 0.$$

This is especially relevant in cases where LC models are used to correct for misclassification (Biemer, 2011), because a violation of an edit restriction is by definition due to misclassification in one of the variables to which the edit restriction applies. By including the edit restriction in the LC model, the appearance of the impossible combination of scores is prevented by constraining the parameter estimates of the LC model.

In the *third step*, m new empty variables are created in the original dataset and imputed by sampling one the LC's using the posterior membership probabilities obtained from the corresponding m LC models:

$$\mathrm{P}(X = x \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = \frac{\mathrm{P}(X = x \mid \mathbf{Z} = \mathbf{z}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x)}{\sum_{x'=1}^{C} \mathrm{P}(X = x' \mid \mathbf{Z} = \mathbf{z}) \prod_{l=1}^{L} \mathrm{P}(Y_l = y_l \mid X = x')}.$$

These posterior membership probabilities represent the probability that a unit is a member of an LC given its combination of scores on the indicators and covariates used in the LC model. At this point, a dataset is obtained containing multiple imputations of the latent variable. From now on, the indicators themselves are no longer needed.

In the *fourth* step, estimates of interest are obtained from the m imputed variables. These can be logistic regression coefficients, tests for model fit, cell proportions in cross tables or any

other estimate of interest to the researcher.

In the *fifth* step, the m estimates are pooled by using the rules defined by Rubin (Rubin, 1987, p.76). The pooled estimate is obtained by:

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i.$$

The total variance is estimated as

$$\text{VAR}_{\text{total}} = \overline{\text{VAR}}_{\text{within}} + \text{VAR}_{\text{between}} + \frac{\text{VAR}_{\text{between}}}{m},$$

where $\overline{\text{VAR}}_{\text{within}}$ is the average within imputation variance and $\text{VAR}_{\text{between}}$ is the between imputation variance. $\overline{\text{VAR}}_{\text{within}}$ is calculated by

$$\overline{\text{VAR}}_{\text{within}} = \frac{1}{m}\sum_{i=1}^{m}\text{VAR}_{\text{within}_i},$$

and $\text{VAR}_{\text{between}}$ is calculated by

$$\text{VAR}_{\text{between}} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})(\hat{\theta}_i - \hat{\theta})'.$$

$\text{VAR}_{\text{between}}$ contains both uncertainty caused by missing or conflicting data and parameter uncertainty (Van der Palm et al., 2016).

5.3.2 Three-step MILC

The MILC method can be expanded to incorporate the three-step approach, enabling the investigation of relationships between latent variable X and auxiliary variables not included in the initial LC model (**Q**). This procedure is shown in Figure 5.2.

The first three steps of the MILC method are applied to create imputations for W: bootstrap samples are created (step *one*), LC models are estimated (step *two*) and m empty variables are imputed (step *three*). An extra step is now required to estimate the classification error of the imputed variables $W_1, .., W_m$ (step *four*):

$$\text{P}(W = w \mid X = x) = \frac{\sum_{\mathbf{y}}\sum_{\mathbf{z}}\text{P}(\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})\text{P}(X = x \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})\text{P}(W = w \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})}{\text{P}(X = x)}.$$

Note that $\text{P}(W = w \mid X = x)$ can be estimated from the imputed dataset directly and that a

separate estimate for $\mathrm{P}(W = w \mid X = x)$ is obtained for every imputation of W.

In step *five*, the relationship between external auxiliary variables $\mathbf{Q}$ and latent variable X ($\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})$) is estimated by using the m imputations obtained in step **three** and the corresponding classification errors obtained in step *four*. $\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})$ is estimated either by using the ML approach (Vermunt, 2010) or the Bolck-Croon-Hagenaars (BCH) approach (Bolck et al., 2004). For both approaches, an LC model is estimated. With the ML approach this is done using a procedure that is comparable to estimating a regular LC model, while for the BCH approach this is done by using a weighting procedure.

With the ML approach, an LC model is specified where W is used as the only indicator of X and this relationship is fixed to the classification error $\mathrm{P}(W = w \mid X = x)$. The form of the $\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})$ distribution is specified in its most general form, and is estimated as:

$$\mathrm{P}(W = w \mid \mathbf{Q} = \mathbf{q}) = \sum_{x=1}^{C} \mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})\mathrm{P}(W = w \mid X = x). \tag{5.1}$$

Based on Equation 5.1, posterior membership probabilities can be obtained for every combination of scores on W and $\mathbf{Q}$:

$$\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q}, W = w) = \frac{\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})\mathrm{P}(W = w \mid X = x)}{\sum_{x'=1}^{C} \mathrm{P}(X = x' \mid \mathbf{Q} = \mathbf{q})\mathrm{P}(W = w \mid X = x')}. \tag{5.2}$$

With the BCH correction method, $\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})$ is estimated by weighting $\mathrm{P}(W = w \mid \mathbf{Q} = \mathbf{q})$ by the inverse of $\mathrm{P}(W = w \mid X = x)$:

$$\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q}) = \sum_{w=1}^{C} P(W = w \mid \mathbf{Q} = \mathbf{q}) d^{*}_{wx},$$

where d^{*}_{wx} represents an element of the inverted $C \times C$ matrix $\mathbf{D}$ with elements $\mathrm{P}(W = w|X = x)$. The obtained result can be plugged into Equation 5.1 to obtain posterior membership probabilities for every combination of scores on W and $\mathbf{Q}$, as shown in Equation 5.2.

The original BCH method has two major drawbacks. First, it can create negative values for the elements $\mathrm{P}(X = x \mid \mathbf{Q} = \mathbf{q})$, resulting in inadmissible solutions. Second, edit restrictions (to prevent the appearance of impossible combinations of scores on X and $\mathbf{Q}$) cannot be incorporated. To circumvent these issues, Chapter 6 placed the BCH approach in a framework of quadratic loss functions and linear equality and inequality constraints. This approach is used throughout the remainder of this paper.

Both the ML and the BCH correction procedure result in a set of posterior membership probabilities for every combination of scores on W and $\mathbf{Q}$ (and for each of the m bootstrap samples), which can be used to create m new imputations for W. The same procedure as followed in step three is used here, although the posteriors are now also conditional on $\mathbf{Q}$. Performing these new imputations is the *sixth* step of the procedure. Step *seven* is then to obtain estimates of interest for each bootstrap sample, which are likely to be parameter estimates describing the relationship between the imputed latent variable W and the external auxiliary variables $\mathbf{Q}$. In step *eight*, these estimates are pooled using Rubin's rules.

5.4 Simulation

5.4.1 Simulation setup

To empirically evaluate the performance of the three-step MILC method, we conducted a simulation study using R (R Core Team, 2014). We started by creating a theoretical population using Latent GOLD (Vermunt & Magidson, 2013a) containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) of the property of interest (X) and two dichotomous variables Q_1 and Q_2 that we consider as external auxiliary variables, so they are not included in the initial LC model. Variations are made according to scenarios described in the subsequent subsections. A theoretical population is used to draw $1,000$ samples and these are used to evaluate the performance of the three-step MILC approach, following the steps described in Section 5.2. In the initial LC model (step *two*), only the three indicators are included in the LC model. At step *five*, the three-step procedure is applied for external auxiliary variables Q_1 and Q_2 simultaneously.

In this simulation study, the performance of two different approaches to the three-step MILC method are evaluated: ML and BCH. As a reference, we also include estimates obtained when no correction method was applied, so where W is imputed using an LC model containing only indicators, and its relationship with Q_1 and Q_2 is investigated directly.

When evaluating the correction methods, the relationship between X and Q_1 and Q_2 should be preserved. There are two types of relationships we are specifically interested in. For the first, we compare the logit coefficient of latent variable X regressed on Q_1 in the theoretical population with the logit coefficient of imputed W regressed on Q_1. This relationship is investigated using four performance measures:

- The bias of the logit coefficient, which is equal to the difference between the average

estimate over all replications and the value found in the theoretical population.

- The coverage of the 95% confidence interval.
- The ratio of the average standard error of the estimate over the standard deviation of the $1,000$ replication estimates is examined to confirm that the standard errors of the estimates are properly estimated.
- The root mean squared error, which is the root of the average of the squares of the errors.

Second, we are interested in a restricted relationship as described in Equation 5.3.1. In the theoretical population, $\mathrm{P}(X = 1 \mid Q_2 = 2) = 0$. When an impossible combination of scores between X and an external auxiliary variable exists but is not accounted for in the LC model, it can appear in the imputed dataset because LC models do not assign probabilities of exactly zero by default. Therefore, we investigate whether the restricted cell in the imputed dataset $\mathrm{P}(W = 1 \mid Q_2 = 2)$ indeed contains zero observations. This is done by investigating the observed frequency of this specific cell (the observed cell proportion is multiplied by the sample size to obtain the observed frequency).

Previous research has shown that the performance of the MILC method is strongly related to the entropy R^2 value of the LC model (Chapter 2). The entropy R^2 indicates how well one can predict class membership based on the observed variables, and is influenced by the measurement quality of the indicators. Therefore, we investigate a range of realistic values for the measurement quality of the indicators in the simulation study. The conclusion was also made that five imputations are sufficient to obtain unbiased estimates and appropriate coverage of the 95% confidence interval (Chapter 2), so $m = 5$ is used in this simulation study as well. Furthermore, different sample sizes are investigated, since they influence the standard errors and thereby the confidence intervals. The main properties of this simulation study are summarized as follows:

- Class-specific response probabilities of the three dichotomous indicators of dichotomous X (Y_1, Y_2, Y_3): 0.70; 0.80; 0.90; 0.95; 0.99 (corresponding entropy R^2 values respectively: 0.31; 0.59; 0.86; 0.96; 0.99).
- Logit coefficients of X regressed on Q_1: 0.00; 0.50; 1.00; 2.00.
- Different proportions for $\mathrm{P}(Q_2 = 2)$, where $\mathrm{P}(W = 1|Q_2 = 2)$ should contain zero observations: $\mathrm{P}(Q_2)$= 0.01; 0.05; 0.10; 0.20.
- Sample size: 200; 500; $1,000$.

- Number of bootstrap samples and multiple imputations $m = 5$.
- Correction methods: No correction method; ML; BCH.

5.4.2 Simulation results

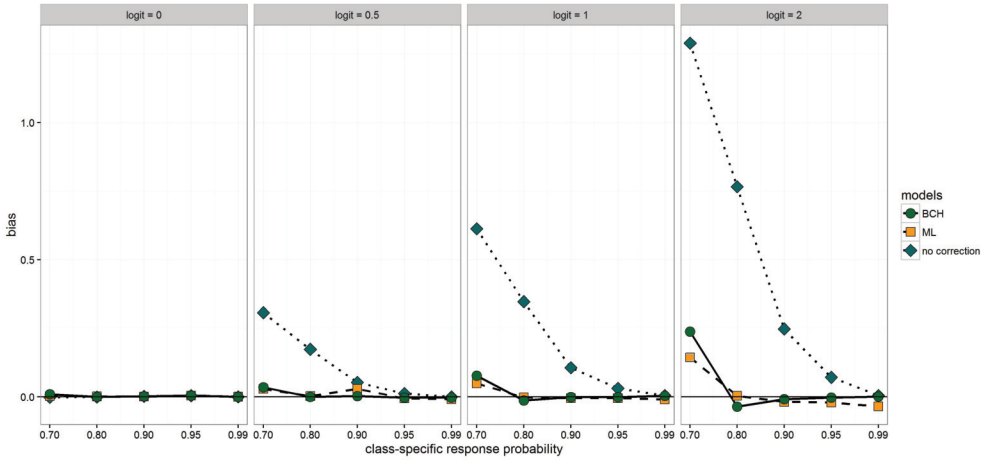

Figure 5.3: Displayed is the bias of the logit coefficient of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variable. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

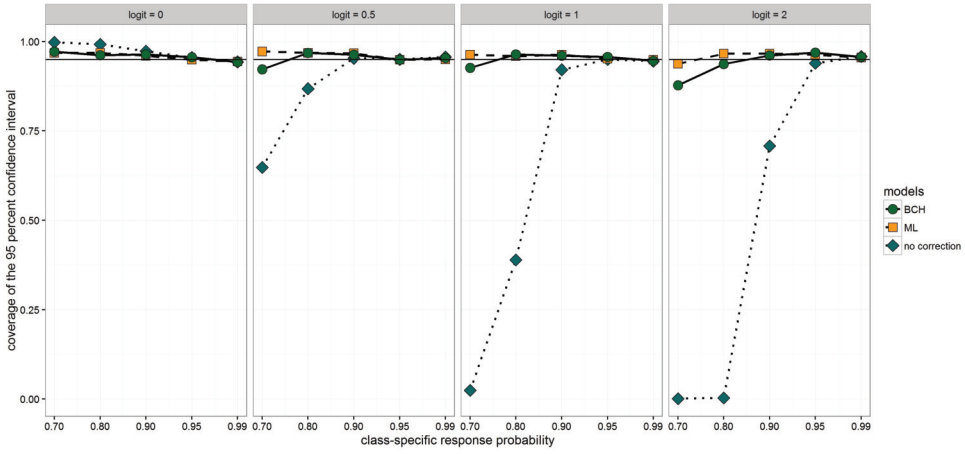

Figure 5.4: Displayed is the coverage of the 95% confidence interval of the logit coefficient of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variable. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

Figure 5.3 shows the bias of the logit coefficient of latent variable X regressed on covariate Q_1

when estimated by using imputed variable W regressed on covariate Q_1. When comparing the results over different strengths of the logit coefficients, in general there is more bias when the logit coefficient increases. When the logit coefficient is zero (i.e. there is no effect), there is approximately no bias in all conditions for both correction methods, and when no correction method is applied. When the logit coefficient increases, bias increases as well if no correction method is applied, while it remains low when correction methods are applied. The only exception is when the class-specific response probabilities are low (0.70).

Figure 5.4 shows the coverage of the 95% confidence interval of the logit coefficient of imputed variable W regressed on covariate Q_1. If the population logit coefficient is zero, the correction methods perform approximately equally well, and not using a correction method also leads to desired results. As the population logit coefficient becomes larger, undercoverage becomes more of a problem when no correction method is applied, especially when the class-specific response probabilities are low. The results obtained for the ML and BCH method are very similar. The coverage rates are generally somewhat higher for ML when the class-specific response probabilities are lower, while the coverage rates are almost identical when the class-specific response probabilities are higher. This is unrelated to the strength of the population logit coefficient.

Figure 5.5 shows the ratio of the average standard error of the estimated logit coefficients over the standard deviation of the logit coefficient of imputed variable W regressed on covariate Q_1. Here we investigate whether the estimated standard errors are indeed equal to the standard deviation of the estimates. When no correction is applied, the standard errors are too large when the class-specific response probabilities are low, and the ratio comes closer to one as the class-specific response probabilities increase. This is unrelated to the size of the population logit coefficient. A comparable trend is seen for the ML and BCH correction methods. The ratio is however much closer to the desired value of one for both methods compared to when no correction method is applied. The trend for BCH method becomes a bit more unstable as the size of the logit coefficient increases, while the trend for the ML method seems more stable.

Figure 5.6 shows us the root mean square error, where the errors are represented by the difference between the logit coefficient of imputed W and Q_1 and its value in the theoretical population. When the logit coefficient is zero, using no correction is the best option in terms of RMSE. However, as soon as the logit coefficient increases, the RMSE of using no correction becomes larger compared to both correction methods. This effect becomes stronger as the logit

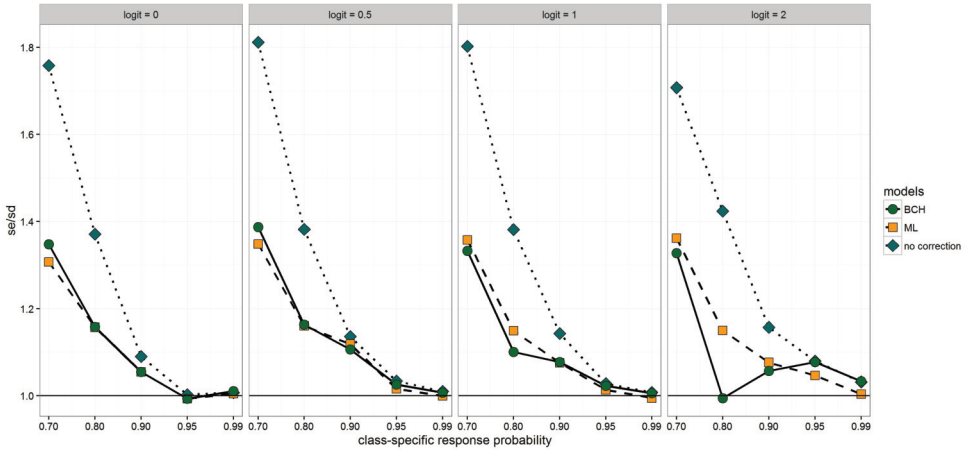

Figure 5.5: Displayed is the ratio of the average standard error of the logit coefficients over the standard deviation of the logit coefficients of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variables. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

coefficient increases. The correction methods perform approximately equally well, where the RMSE is generally a bit lower for the ML method.

Figure 5.7 shows the number of times that the combination of scores that is in practice impossible, $\mathrm{P}(W = 1 \mid Q_2 = 2)$, is observed in the imputed dataset averaged over $1,000$ replicates. We see the results when no correction is applied, and for the ML and BCH correction methods. Furthermore, we see the results for different class-specific response probabilities. When no correction is applied, the observed frequency (the cell proportion multiplied with the sample size) is strongly related to these class-specific response probabilities. We see that when the class-specific response probabilities are low (0.70), the observed frequency of $\mathrm{P}(W = 1 \mid Q_2 = 2)$ in this condition is around 65. This number decreases as the class-specific response increases, but even when the class-specific response probabilities are 0.90, there are still impossible combinations of scores created. When the ML correction method is applied, zero impossible combinations of scores are created under all conditions investigated, only when the class-specific response probabilities are 0.70, the number of impossible combinations created is not exactly zero, but still below one. With the BCH correction method, impossible combinations of scores are only created when the class-specific response probabilities are 0.70. This makes sense, the entropy R^2 is very low in these conditions so we did not expect the correction methods to perform well here. In all other

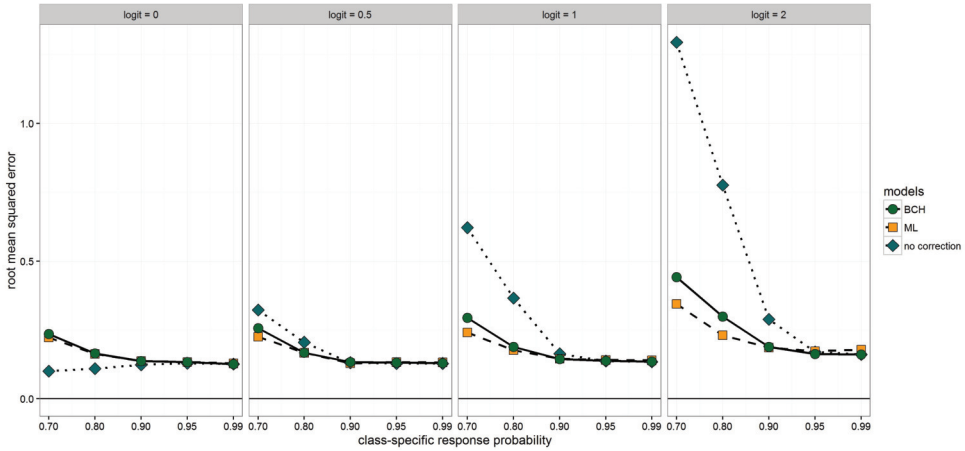

Figure 5.6: Displayed is the root mean squared error, where the errors are represented by the difference between the logit coefficient of imputed W on Q_1 and its value in the theoretical population. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variables. Sample size is 1,000 and $\mathrm{P}(Z = 2) = 0.2$.

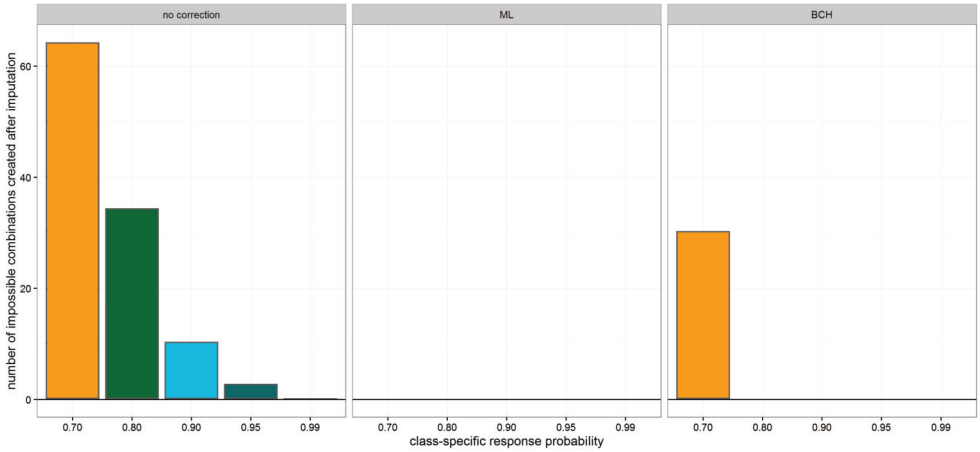

Figure 5.7: The bars in this histogram display the number of times that the impossible combination $\mathrm{P}(W = 1 \mid Q_2 = 2)$ is observed when no correction, ML and BCH are applied. The sample size is $1,000$ and the marginal of Q_2 is 0.2. Results are displayed for different class-specific response probabilities of the indicator variables.

conditions, no impossible combinations of scores are created.

Overall it can be said that problems in terms of bias and coverage of estimates can be severe if no correction is applied when performing a latent class three-step method. Both correction methods (ML and BCH) have shown to improve these results and the differences in

results between the methods are minimal. Even with low class-specific response probabilities, improvements were detected for both methods, although not all problems are solved in these low-quality cases. For example, BCH was not able to handle edit restrictions in combination with low class-specific response probabilities and the coverage rates for BCH were also somewhat lower in these cases.

5.5 Applications

In this section, the flexibility of the three-step MILC method is illustrated by applying the method to two empirical datasets. First, the method is applied to a composite dataset used in official statistics, where researchers use LC models to correct for measurement error. Second, the method is applied to a dataset containing scores on math items, where discretized IRT models can be used to measure mathematical ability.

In each application, the relationship between an imputed latent variable and an external auxiliary variable is investigated. First, an LC model is applied without including the auxiliary variable, this will be denoted as uncorrected MILC. The relationship between the imputed latent variable and the auxiliary variable is then investigated. Next, the estimate of this relationship is corrected using the ML and BCH method. For comparison, the estimate of this relationship is also investigated when the external variable was included in the initial LC model.

5.5.1 A latent class model to correct for measurement error in official statistics

We investigate the relationship between *home ownership* and *marital status*. To estimate this relationship, a composite dataset is used that consists of two surveys carried out by LISS (Longitudinal Internet Studies for the Social sciences) from 2013 (Scherpenzeel, 2011), which is administered by CentERdata (Tilburg University, The Netherlands) and a population register from Statistics Netherlands from 2013. Since this composite dataset contains two variables indicating whether a person is either a 'home-owner' or 'home-renter or other', we use these as indicators to measure the *true* variable 'home-owner'/'home-renter or other', which can correct for misclassification in the indicator variables. Since an LC model with only two indicators is not identifiable, we also included a covariate measuring whether someone receives *rent benefit* from the government. Since a person can only receive rent benefit if this person rents a house, we included an edit restriction here. For a detailed description of the combined dataset and the processing it, we refer to Chapter 2. Next, we impute

the *true* variable measuring 'home-owner'/'home-renter or other' using the LC model, and we investigate the relationship between this variable and a covariate measuring *marital status*. More specifically, we investigate whether *marriage* can predict *home ownership*. First,

Table 5.1: The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the variable *owning/renting a house*. The first row represents the results obtained when no correction method is applied. The second and third row represent the ML and BCH correction methods. The last row represents the results obtained when the auxiliary variable *marital status* is included in the initial LC model.

	intercept		marriage	
	estimate	95% CI	estimate	95% CI
No correction	-2.6829	[-2.9533; -2.4126]	1.2524	[0.9820; 1.5227]
ML	-2.7221	[-2.9898; -2.4544]	1.3247	[1.0570; 1.5924]
BCH	-2.6097	[-2.8669; -2.3526]	1.3866	[1.1294; 1.6437]
Included	-2.7712	[-3.0389; -2.5036]	1.3817	[1.1140; 1.6493]

uncorrected MILC was applied (*marriage* was not included in the LC model). Second, MILC was applied while the auxiliary variable *marriage* was included as a covariate in the LC model. Results of both these models can be found in Table 5.1. Here we see that both the intercept and the logit coefficient are closer to zero when the auxiliary variable was not included in the initial model, compared to when it is included.

Furthermore, if we apply either the ML or BCH method to correct the imputations made using a model without covariate, we see that the differences between these corrected estimates and the estimates when including the auxiliary variable into the model are much smaller, although they are not exactly equal.

In general we can conclude that non-married individuals are approximately equally likely to own than rent a house (non-married individuals are approximately $e^{-2.7} = 0.07$ times more likely to own than rent a house). Married individuals are more likely to own a house than to rent it. However, if we would not have included this auxiliary variable in the LC model, we would conclude that they are approximately $e^{1.2524} = 3.4987$ times more likely to own than to rent a house. If we would have either included the auxiliary variable in the model or used a correction method, we would conclude that this relation is actually somewhat stronger, either $e^{1.3247} = 3.7611$ (ML), $e^{1.3866} = 4.0012$ (BCH) or $e^{1.3817} = 3.9817$ (included) times more likely. In general, the results from the included model and from both the correction methods are quite close.

5.5.2 A discretized IRT model in psychometrics

Here we investigate the relationship between *mathematical ability* of a child and the *level of education* of its mother. To investigate this relationship, we make use of the 2015 PISA data. More specifically, we use a subset of the data containing 141 Dutch 15-year old pupils who conducted booklet number 43 of the mathematical ability test. This booklet contains 20 mathematical ability questions which can be graded with either 'correct'/'incorrect' or 'correct'/'partly correct'/'incorrect'. The scores obtained by answering these questions are used as indicators of the latent variable *mathematical ability*, which we measure using a discretized IRT model in Latent GOLD with two classes ('low level'/'high level'). We impute the latent variable of *mathematical ability* using the discretized IRT model, and we investigate the relationship between *mathematical ability* of the child and its mothers' *level of education*. To measure the *level of education* of the mother, we used a dichotomous variable indicating whether she finished the International Standard Classification Level 4 (ISCL 4: post-secondary non-tertiary education). We investigate whether mothers' *level of education* can predict *math*

Table 5.2: The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the outcome variable *math ability*. The first row represents the results obtained when no correction method is applied. The second and the third row represents the results of the ML and BCH correction methods, the last row represents the results obtained when the auxiliary variable is included in the initial LC model.

	intercept		high ability	
	estimate	95% CI	estimate	95% CI
No correction	0.6837	[0.1917; 1.1757]	-1.0867	[-1.8377; -0.3358]
ML	0.7259	[0.2223; 1.2294]	-1.1329	[-1.8794; -0.3864]
BCH	0.7351	[0.2176; 1.2527]	-1.0726	[-1.8128; -0.3325]
Included	0.7421	[0.1767; 1.3075]	-0.9293	[-1.6718; -0.1869]

ability of her child. First, uncorrected MILC was applied (where mothers' *level of education* was not included in the LC model as a covariate). Here, the intercept is $e^{0.6837} = 1.9812$, which can be interpreted as the odds that a mother has *education level* ISCL 4 when a child has *mathematical ability* 'above level'. The logit coefficient is $e^{-1.0867} = 0.3373$. The odds for a mother to have *education level* ISCL 4 when her child does not have *mathematical ability* 'above level' is 0.3373 times the odds when her child has *mathematical ability* 'below level'.

In this example, it can be particularly undesirable that mothers' *level of education* is included in the LC model as a covariate, since it then contributes to the assignment of children to classes measuring their *mathematical ability*, and researchers can find it undesirable that children are assigned to a *math ability* class based on their mothers' ability. It can be seen in the results that the relationship between the two variables is stronger when the mothers' *level of education* is

included in the model. When MILC is applied with mothers' *level of education* included in the LC model as covariate, this logit coefficient is $e^{-0.9293} = 0.3948$, so the estimated relationship between mothers' *level of education* and *math ability* of her child is stronger when this variable is included in the model compared to when it's not included in the model.

However, when the model without covariate is used, there is no correction for the fact that the relationship of interest is estimated using an imputed version of *math ability* and not the true values of *math ability*. Therefore, correction methods are applied and they both result in small adjustments compared the uncorrected results. For both ML and BCH, the intercept is a bit larger compared to the uncorrected results, while the logit coefficient is a bit smaller.

5.6 Discussion

In this paper we introduced the three-step MILC method, which updates latent variable imputations to be conditional on external variables. If the latent variable imputation is not corrected to be conditional on external variables, the point estimates of the relationship between the imputed latent variable and the external variables are biased. This bias is caused by the fact that an imputation of a latent variable is generally not a perfect representation of that latent variable, it contains some measurement error. While a method that corrects for classification error is required, current methods lack either general applicability or flexibility. Therefore, the MILC method and the three-step approach of the latent class model are combined into one generic procedure.

We incorporated two alternative correction procedures in the three-step MILC method, and evaluated them in terms of their ability to correct for bias in point estimates due to measurement error in the imputed latent variable. We assessed the different procedures in terms of bias of the estimates, coverage of their 95% confidence interval, standard error of the estimate over the standard deviation over the estimates and root mean squared error. Furthermore, we investigated whether the different correction procedures were able to successfully incorporate edit restrictions. This all was investigated under a number of different conditions in a simulation study.

From the simulation study, it can be concluded that the necessity of applying the three-step MILC method (or probably any other correction procedure) was strongly related to the strength of the relationship under investigation. If the true logit coefficient was zero, i.e. there was no relationship between the imputed latent variable and the external covariate, then there

was also no bias if no correction procedure was used. In other words, there was nothing to correct. Furthermore, the necessity of applying the three-step MILC was also related to class separation.

When class separation was higher, results of better quality were obtained when no correction was applied. However, regardless of the strength of the relationship under investigation or the strength of the class separation, results always improved when a correction method was applied compared to when no correction method was applied. Furthermore, the BCH and ML correction methods performed in a very comparable way. ML can be recommended over BCH in cases where the class-specific response probabilities of the indicators are low, since the simulation results showed that the coverage rates where somewhat lower for BCH here, and they showed that BCH was not able to succesfully incorporate edit restrictions in these situations.

It should also be noted that estimates of interest can also be obtained directly after applying the BCH or ML correction procedure. These results are then obtained from the output generated by these correction procedures and not by creating new imputations and investigating these. To be able to directly obtain these corrected estimates, the researcher needs to think about the type of relationship that he or she wants to investigate and specify the correction procedure correspondingly. However, using an imputed variable allows for much more flexibility, because when investigating this variable in relationship with other variables, the researcher is not limited to how these relationships are specified in the correction procedure.

The two applications show the great flexibility of the three-step MILC method. In the first application, a composite dataset from official statistics is used, where LC models are applied to correct for classification error. In the second application, a dataset containing children's scores on math items is used, where a discretized IRT model is used to investigate children's *math ability*. In both applications, both ML and BCH perform approximately equally well. Furthermore, the second application also shows that directly including an external variable into the LC model can have an undesirable influence on the class assignments.

The results in this paper invite further investigation of the applicability of the three-step MILC method to continuous data or other types of models that can be fit into the latent class framework, such as latent Markov models or multilevel latent class models.

Further research into the three-step MILC method is also required due to the limited scope of the current simulation study. Only a small latent class model is investigated, where various model assumptions were made. For example, the auxiliary variables were assumed to be free of measurement error. This can be strange in an official statistics setting where the LC model itself is used to correct for measurement error in the indicator variables. In addition, the indicator variables are assumed to be conditionally independent and the classification error in these indicators is assumed to be unrelated to the auxiliary variables. Unfortunately, these assumptions will not always be met in practice and thus how robust this method is to violations of these assumptions should be looked into.

In summary, the three-step extension of the MILC method presented in this paper allows correct estimation of relationships between an imputed latent variable and external auxiliary variables. This method is a promising solution to correct for misclassification, due to its general applicability and flexibility.

Chapter 6

A note on applying the BCH method under linear equality and inequality constraints

Abstract

Researchers often wish to relate estimated scores on latent variables to exogenous covariates not previously used in analyses. The BCH method corrects for asymptotic bias in estimates due to these scores' uncertainty and has been shown to be relatively robust. When applying the BCH approach however, two problems arise. First, negative cell proportions can be obtained. Second, the approach cannot deal with situations where marginals need to be fixed to specific values, such as edit restrictions. The BCH approach can handle these problems when placed in a framework of quadratic loss functions and linear equality and inequality constraints. This research note gives the explicit form for equality constraints and demonstrates how solutions for inequality constraints may be obtained using numerical methods.

This chapter was published as: L. Boeschoten, M.A. Croon, D.L. Oberski (2018) A note on applying the BCH method under linear equality and inequality constraints *Journal of Classification*, 1 - 10

6.1 Introduction

Researchers in many different disciplines apply latent structure models in which observed variables are treated as indicators of an underlying latent variable that can not be measured directly. An often used strategy in this context consists of three steps (Vermunt, 2010). First, the parameters of the measurement model are estimated, describing the relationship between the latent variable and its indicators. Second, each respondent is assigned a latent score based on his/her scores on the indicators. Finally, the relationships between the latent scores and scores on exogenous variables are assessed.

Croon et al. (2002) showed that for general latent structure models such a strategy leads to inconsistent estimates of the parameters of the joint distribution of the latent variable and the exogenous variables. Bolck et al. (2004) discussed this problem in the context of latent class analysis where observed variables are categorical. They also derived a correction procedure that produces consistent estimates, known as the BCH correction method. Subsequent simulation studies by Vermunt (2010), Bakk et al. (2013), Bakk & Vermunt (2016) and Nylund-Gibson & Masyn (2016) have demonstrated that this procedure produces unbiased parameter estimates and correct inference for a large range of simulation conditions. When applying the BCH correction method in cases of categorical exogenous variables, two problems can arise. First, negative cell proportion estimates can be obtained (Asparouhov & Muthén, 2014). Second, the approach cannot deal with situations where marginals need to be constrained. An example is edit restrictions in official statistics, leading to certain marginals being fixed to zero (De Waal et al., 2011), which is also used in combination with latent class modelling (Chapter 2).

In this research note the BCH method is extended to solve these two problems. We allow for linear equality and inequality constraints by noting the correction method minimizes a quadratic loss function and give a closed form solution for linear equality restrictions. Next, we demonstrate how solutions for inequality constraints may be obtained using numerical methods. We first discuss the three-step approach to the latent class model and the BCH correction method. We then show how to impose linear restrictions and how to extend this to including non-negativity constraints. At last, the extended BCH method is applied on a dataset from the Political Action Survey. In the Appendix E, R code is given to apply the procedure.

6.2 The three-step approach to the latent class model and the BCH correction method

Let us denote a set of observed exogenous variables $\mathbf{Q}$ and an unobserved latent variable X. All variables involved are assumed to be categorical. Let $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_J)$ be the Cartesian product of J different discrete random variables $\mathbf{Q}_j$. If the variable $\mathbf{Q}_j$ is defined for n_j categories, the distribution of $\mathbf{Q}$ can be specified as a multinomial distribution with $n = \prod_{j=1}^{J} n_j$ categories.

In the basic latent class model considered by Bolck et al. (2004), a single categorical latent variable X with m categories is introduced. The variable X itself is not directly observed but only indirectly via a set of indicator variables $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_K)$. Let the joint distribution of the categorical variables $\mathbf{Q}$, X and $\mathbf{Y}$ be denoted by

$$\mathrm{p}(\mathbf{Q} = \mathbf{q}, X = x, \mathbf{Y} = \mathbf{y}) = \mathrm{p}(\mathbf{q}, x, \mathbf{y}).$$

Then a possible factorization is

$$\mathrm{p}(\mathbf{q}, x, \mathbf{y}) = \mathrm{p}(\mathbf{q})\mathrm{p}(x|\mathbf{q})\mathrm{p}(\mathbf{y}|x, \mathbf{q}).$$

Since in the basic latent class model $\mathbf{Q}$ is assumed to have no direct effect on $\mathbf{Y}$, the latter result simplifies to

$$\mathrm{p}(\mathbf{q}, x, \mathbf{y}) = \mathrm{p}(\mathbf{q})\mathrm{p}(x|\mathbf{q})\mathrm{p}(\mathbf{y}|x).$$

The three-step approach to the estimation of the parameters of the latent class model starts with the estimation of the parameters of the measurement model represented by the conditional probability distribution $\mathrm{p}(\mathbf{y}|x)$. Once this estimation procedure is completed, individual research units may be assigned to one of the latent classes solely on the basis of their observed scores on $\mathbf{Y}$. This defines the second step of the estimation procedure and results in an assignment of each individual to a latent class. If the random variable W represents the latent classes individuals are assigned to, and assignment is done using a modal rule where each individual is assigned to the class for which its posterior membership probability is the largest, this can be expressed as

$$\mathrm{p}(w|\mathbf{y}) = \begin{cases} 1 & \text{if } \mathrm{p}(x_1|\mathbf{y}) > \mathrm{p}(x_2|\mathbf{y}) \; \forall \; x_1 \neq x_2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Different assignment rules than the modal rule will yield a different form for Equation 6.1.

All subsequent results also apply to other assignment rules, such as proportional or random assignment (Bakk, 2015).

Since $\mathbf{Y}$ and $\mathbf{Q}$ are conditionally independent given X, so are W and $\mathbf{Q}$ and the conditional distributions are related by

$$\mathrm{p}(w|\mathbf{q}) = \sum_{x=1}^{\mathbf{X}} \mathrm{p}(w|x)\mathrm{p}(x|\mathbf{q}).$$

In terms of the joint distribution this becomes

$$\mathrm{p}(\mathbf{q}, w) = \sum_{x=1}^{\mathbf{X}} \mathrm{p}(\mathbf{q}, x)\mathrm{p}(w|x).$$

The latter result can be recast as a matrix equation

$$\mathbf{E} = \mathbf{AD},$$

with the elements of the three matrices defined as $e_{\mathbf{q}w} = \mathrm{p}(\mathbf{q}, w)$, $a_{\mathbf{q}x} = \mathrm{p}(\mathbf{q}, x)$ and $d_{xw} = \mathrm{p}(w|x)$. After completing the first and the second estimation step, the elements of the matrices $\mathbf{E}$ and $\mathbf{D}$ are known. The joint distribution of $\mathbf{Q}$ and the latent variable X is then given by

$$\mathbf{A} = \mathbf{E}\mathbf{D}^{-1}.$$

Here it is assumed that matrix $\mathbf{D}$ is not singular so that its inverse exists. See (Bolck et al., 2004, pp.13-14) for a discussion on when this assumption may be violated. A consistent estimate of $\mathbf{A}$ is $\hat{\mathbf{E}}\hat{\mathbf{D}}^{-1}$.

The previously obtained algebraic solution for matrix $\mathbf{A}$ can also be derived via a rather trivial minimization of a least squares function. Let $\mathbf{E}$ and $\mathbf{D}$ be matrices with known elements. Matrix $\mathbf{E}$ is of order $n \times m$ and $\mathbf{D}$ is an invertible matrix of order $m \times m$. Let $\mathbf{A}$ be an $n \times m$ matrix of unknown elements and consider the following least squares function:

$$\varphi = \frac{1}{2}\mathrm{tr}(\mathbf{AD} - \mathbf{E})'(\mathbf{AD} - \mathbf{E}).$$

Minimizing φ with respect to the unknown matrix $\mathbf{A}$ yields $\mathbf{A} = \mathbf{E}\mathbf{D}^{-1}$, for which φ attains the truly minimal value of zero. Note that the factor $1/2$ is introduced to obtain simpler expressions for the first derivatives. Its introduction does not change the solution of the minimization problem.

6.3 The correction procedure under linear equality constraints

In some applications, simple linear restrictions may be imposed on the elements of matrix **A**. For instance, some of the probabilities in the joint distribution of **Q** and X may be set equal to zero, for example for combinations of **Q** and X that cannot occur in practice. After imposing such zero constraints, all the non-zero cell probabilities should still add to one. The quadratic loss function φ can be minimized under equality constraints on the unknown elements of matrix **A** by applying the method of Lagrangian multipliers.

We first rewrite the quadratic loss function φ in the following way using vectorization operations on matrices, see Schott (1997, p. 261-266). For the vector of residuals $\mathbf{r}$ we obtain

$$\begin{aligned}\mathbf{r} &= \text{vec}(\mathbf{AD} - \mathbf{E}) \\ &= \text{vec}(\mathbf{I}_{n\times n}\mathbf{AD}) - \text{vec}(\mathbf{E}),\end{aligned}$$

where $\mathbf{I}_{n\times n}$ is an $n \times n$ identity matrix. Applying Theorem 7.15 from Schott (1997, p. 263) yields

$$\mathbf{r} = (\mathbf{D}' \otimes \mathbf{I}_{n\times n}) \cdot \text{vec}(\mathbf{A}) - \text{vec}(\mathbf{E}),$$

in which $\otimes$ is the Kronecker product of two matrices (A. Graham, 1982). Defining $\mathbf{P} = \mathbf{D}' \otimes \mathbf{I}_{n\times n}$, $\mathbf{a} = \text{vec}(\mathbf{A})$ and $\mathbf{e} = \text{vec}(\mathbf{E})$, we are able to write

$$\mathbf{r} = \mathbf{Pa} - \mathbf{e},$$

so that the least squares function becomes

$$\begin{aligned}\varphi &= \frac{1}{2}\mathbf{r}'\mathbf{r} \\ &= \frac{1}{2}(\mathbf{a}'\mathbf{P}'\mathbf{Pa} - 2\mathbf{e}'\mathbf{Pa} + \mathbf{e}'\mathbf{e}).\end{aligned}$$

The completely unconstrained solution to the minimization problem is given by

$$\mathbf{a}_0 = (\mathbf{P}'\mathbf{P})^{-1} \cdot \mathbf{P}'\mathbf{e}.$$

Now suppose that the S linear equality constraints can be represented by a matrix equation

$$\mathbf{Ha} = \mathbf{c}.$$

The matrix $\mathbf{H}$ is of order $S \times N$, N being the number of cells in matrices **A** and **E**. We may assume that $\mathbf{H}$ is of rank S; otherwise, the linear equality constraints would not be

linearly independent. In order to minimize the least square function φ under a set of S linear constraints on the elements of $\mathbf{A}$, the Lagrangian is defined as

$$\mathbf{L} = \varphi - \lambda'(\mathbf{Ha} - \mathbf{c}). \tag{6.2}$$

Setting the first derivatives of $\mathbf{L}$ with respect to $\mathbf{a}$ equal to the zero vector, and solving for $\mathbf{a}$ yields:

$$\mathbf{a} = (\mathbf{P}'\mathbf{P})^{-1}(\mathbf{P}'\mathbf{e} + \mathbf{H}'\lambda),$$

which can be rewritten as:

$$\mathbf{a} = \mathbf{a}_0 + (\mathbf{P}'\mathbf{P})^{-1}\mathbf{H}'\lambda.$$

Solving for the unknown Lagrangian multipliers by taking the derivative of the Lagrangian in Equation 6.2, and setting it to zero, or equivalently by imposing linear constraints $\mathbf{Ha} - \mathbf{c} = 0$ yields:

$$\lambda = [\mathbf{H}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{H}']^{-1}(\mathbf{c} - \mathbf{Ha}_0).$$

So that the final solution for $\mathbf{a}$ is:

$$\mathbf{a} = \mathbf{a}_0 + (\mathbf{P}'\mathbf{P})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{H}']^{-1}(\mathbf{c} - \mathbf{Ha}_0).$$

Note that the vector $\mathbf{c} - \mathbf{Ha}_0$ represents the deviations of the unconstrained solution from the linear equality constraints. Again a consistent estimate of $\mathbf{a}$ can be obtained by replacing $\mathbf{P}$ and $\mathbf{a}_0$ with their sample estimates.

6.4 The correction procedure under linear equality and inequality constraints

A second issue with the BCH procedure is that in finite samples the consistent estimate $\hat{\mathbf{A}}$ hat may contain negative values. This issue is similar to the occurrence of Heywood cases in factor analysis (Heywood, 1931). Such negative values in the probability table estimate $\hat{\mathbf{A}}$ may prevent subsequent analyses. We suggest to prevent such inadmissible solutions by imposing inequality constraints. The resulting minimization problem is a quadratic program that can be solved by an iterative method.

Such a numerical iterative method for an equality and inequality constrained minimization of a quadratic function has been described by Goldfarb & Idnani (1983). Their numerical

algorithm solves the quadratic programming problem of the form

$$\min(\frac{1}{2}\mathbf{b}'\mathbf{D}_{\text{mat}}\mathbf{b} - \mathbf{d}'_{\text{vec}}\mathbf{b}),$$

subject to the constraints

$$\mathbf{H}'\mathbf{b} \geq \mathbf{b}_0,$$

with respect to the n unknown parameters in vector $\mathbf{b}$. The matrix $\mathbf{D}_{\text{mat}}$ is a given $n \times n$ symmetric positive definite matrix whereas $\mathbf{d}_{\text{vec}}$ is a given $n \times 1$ vector.

In order to apply the Goldfarb-Idnani optimization procedure in the present context, the following definitions have to be implemented. First, to include non-negativity constraints, we make use of Theorem 7.6 from Schott (1997, p.254) to obtain

$$\begin{aligned}\mathbf{D}_{\text{mat}} &= \mathbf{P}'\mathbf{P} \\ &= (\mathbf{D}\mathbf{D}') \otimes \mathbf{I}_{n\times n}.\end{aligned}$$

and

$$\mathbf{d}_{\text{vec}} = \mathbf{P}'\mathbf{e}$$

Since it is assumed that matrix $\mathbf{D}$ is of full rank, the matrix $\mathbf{P}'\mathbf{P}$ is positive-definite. This ensures that the quadratic loss function φ is strictly convex. Moreover, the type of equality and inequality constraints considered here (the sum of the elements in matrix $\mathbf{A}$ is equal to 1, where all elements ≥ 0 and some are fixed to zero), define a convex region in the parameter space.

In order to represent the constraints on the cell probabilities we now define matrix $\mathbf{H}$ in such a way that the first row of $\mathbf{H}$ has all its elements equal to one. This row represents a constraint on the sum of all cell probabilities. We represent this row vector as matrix $\mathbf{H}_0$. Let $J = \{1, 2, 3, ..., N\}$ be an index set corresponding to the column numbers of matrix $\mathbf{H}$. This index set can be partitioned in two non-overlapping subsets J_1 and J_2:

- Subset J_1 contains the indices of the elements of vector $\mathbf{a}$ which are set exactly equal to zero: for those indices j we require $\mathbf{a}_j = 0$;
- Subset J_2 contains the indices of the elements of vector $\mathbf{a}$ which are required to be non-negative: for those indices j we require $\mathbf{a}_j \geq 0$.

Now let I_n be an $N \times N$ identity matrix and permute the rows of this matrix so that the upper

part contains the rows corresponding with the index numbers in J_1, and the lower part of the permuted identity matrix contains the rows corresponding with the index numbers in J_2. Referring to the two parts of the permuted identity matrix as $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively, the matrix $\mathbf{H}$ is obtained by

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1 \\ \mathbf{H}_2 \end{pmatrix}$$

where $\mathbf{H}$ is used to obtain the final solution for $\mathbf{a}$. Note that in cases where we are not interested in applying equality constraints, but we are interested in applying the inequality constraints we simply omit $\mathbf{H_1}$. Vector $\mathbf{b_0}$ is of length $N + 1$, with its first element equal to 1 and all the remaining elements equal to 0.

With this procedure, we are able to find a solution for $\mathbf{A}$ (the joint distribution of latent variable X and exogenous covariates $\mathbf{Q}$) where the sum of the elements is equal to 1, where no negative elements are created, and where impossible combinations of scores can be set to have a probability of zero. Having defined $\mathbf{b}$, $\mathbf{D}_{\text{mat}}$ and $\mathbf{H}$, the solution can be obtained using standard software for quadratic programming, such as the R package `quadprog` (Turlach & Weingessel, 2013).

6.5 Application

As an illustration, the extended BCH method is applied on a dataset from the Political Action Survey (Barnes et al., 1979; Jennings et al., 1990) . The dataset consists of five dichotomous indicators on political involvement and tolerance (*System Responsiveness*; *Ideological Level*; *Repression Potential*; *Protest Approval*; *Conventional Participation*) and three nominal covariates (*Sex*; *Level Of Education*; *Age*). This dataset has previously been used in Hagenaars (1993), Vermunt & Magidson (2000) and in the Latent GOLD user's manual (Vermunt & Magidson, 2013a). The dataset as well as the syntax used in this illustration can be found in Latent GOLD version 5.1 under 'syntax examples' → LCA → restrictions → equalities → Model C.

In the first step, a four class restricted model is applied to distinguish between four latent classes on involvement and tolerance. In this model, response probabilities are restricted to be equal for the items *System Responsiveness* and *Conventional Participation*, and the response probability for the variable *Ideological Level* is fixed to zero by specifying a logit of 100.

In the second step, cases are assigned to a latent class by using modal assignment, resulting in the imputed latent variable W. In the third step the relationship between the imputed latent variable *Involvement And Tolerance* (W) and exogenous covariate *Age* (Q) is investigated. The **E**-matrix containing the joint probabilities of these variables is:

$$\mathbf{E} = \begin{array}{c} \\ Q_{\text{16-34}} \\ Q_{\text{35-57}} \\ Q_{\text{58-91}} \end{array} \begin{array}{c} \begin{array}{cccc} W_1 & W_2 & W_3 & W_4 \end{array} \\ \begin{pmatrix} 0.05795848 & 0.15743945 & 0.01643599 & 0.09256055 \\ 0.08477509 & 0.17560554 & 0.05276817 & 0.03979239 \\ 0.12802768 & 0.10034602 & 0.06920415 & 0.02508651 \end{pmatrix} \end{array}.$$

The **D**-matrix describing the relationship between the imputed latent variable *Involvement And Tolerance* (W) and the latent variable *Involvement And Tolerance* (X) is also obtained:

$$\mathbf{D} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} W_1 & W_2 & W_3 & W_4 \end{array} \\ \begin{pmatrix} 0.67389148 & 0.1570985 & 0.02678610 & 0.1422239 \\ 0.01898361 & 0.7891416 & 0.05879905 & 0.1330757 \\ 0.17186997 & 0.2725275 & 0.54176422 & 0.0138383 \\ 0.12184782 & 0.3220914 & 0.01975761 & 0.5363031 \end{pmatrix} \end{array}.$$

The BCH method can now be applied by estimating $\mathbf{ED}^{-1}$, resulting in the **A** matrix:

$$\mathbf{A}_{\text{unconstraint}} = \begin{array}{c} \\ Q_{\text{16-34}} \\ Q_{\text{35-57}} \\ Q_{\text{58-91}} \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0.0577223 & 0.13465976 & 0.008359502 & 0.123652898 \\ 0.1018944 & 0.17635045 & 0.073167182 & 0.001529175 \\ 0.1618782 & 0.06157076 & 0.113576159 & -0.014360760 \end{pmatrix} \end{array}.$$

As can be seen, this result is inadmissable since the cell $Q_{\text{58-91}} \times X_4$ contains a negative value. Therefore, it will not be possible to estimate posterior membership probabilities and to do subsequent analyses here.

When the extended BCH method is applied, the following constrained **A** matrix is obtained:

$$\mathbf{A}_{\text{constraint}} = \begin{array}{c} \\ Q_{\text{16-34}} \\ Q_{\text{35-57}} \\ Q_{\text{58-91}} \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0.05741718 & 0.13472999 & 0.007627791 & 0.1229631559 \\ 0.10158926 & 0.17642067 & 0.072435471 & 0.0008394325 \\ 0.15689781 & 0.05436459 & 0.114714655 & 0.0000000000 \end{pmatrix} \end{array}.$$

The cell $Q_{58\text{-}91} \times X_4$ does not contain a negative value anymore, so this matrix can now be used to estimate posterior membership probabilities and to do subsequent analyses.

Since there are no combinations of scores between *Involvement And Tolerance* and *Age* that are not possible in practice, it is not needed to fix any marginals to zero.

6.6 Conclusion

We have modified the BCH method to include linear equality and inequality constraints solving the problem of negative solutions and allowing for restrictions on arbitrary cell margins. With these adjustments, analysts interested in relating covariates to assignments on latent class variables will now be able to, for example, impose edit restrictions, further analyse solutions that were previously inadmissible and analyse datasets involving more complex marginal restrictions. The application demonstrates that when a negative value is obtained using the regular BCH method, this can be solved by using the extended BCH method. In the Appendix, R code is given to apply the extended BCH method, and an addition to the example is given that demonstates how margins can be fixed to zero using the extended BCH method.

PART IV

Using the MILC method for census purposes

Chapter 7

Using Multiple Imputation of Latent Classes (MILC) to construct population census tables with data from multiple sources

Abstract

The MILC method combines multiple imputation and latent class analysis to correct for misclassification in combined datasets. Furthermore, MILC generates a multiply imputed dataset which can be used to estimate different statistics in a straightforward manner, ensuring that uncertainty due to misclassification is incorporated when estimating the total variance. In this paper, it is investigated how the MILC method can be adjusted to be applied for census purposes. More specifically, it is investigated how the MILC method deals with a finite and complete population registry, how the MILC method can simultaneously correct misclassification in multiple latent variables and how multiple edit restrictions can be incorporated. A simulation study shows that the MILC method is in general able to reproduce cell frequencies in both low- and high-dimensional tables with low amounts of bias. In addition, variance can also be estimated appropriately, although more variance is estimated than necessary when cell frequencies are small.

7.1 Introduction

Population and housing censuses provide a picture about the socio-demographic and socio-economic situation of a country and it is ubiquitous that a census should cover the entire population of people and dwellings that are present in a country. Every ten years the United Nations Economic and Social Council (ECOSOC) adopts a resolution, urging Member States to carry out a population and housing census and to disseminate census results as an essential source of information, see e.g. The Economic and Social Council (2005). In the EU, explicit agreements have been made about which variables should be listed in the census, and also which cross-tables between which variables should be produced (European Commission, 2008, 2009 and 2010).

The vast majority of countries produce census data by conducting a traditional census, which entails interviewing inhabitants in a complete enumeration, reaching every single household. Statistics Netherlands can rely on population registries as a main source for most census tables because these registries are of relatively good quality, including a very broad coverage (Geerdinck et al., 2014). All register variables are available from Statistics Netherlands' system of social statistical datasets (B. F. Bakker et al., 2014). The backbone is the Central Population Register which combines the population registries from the municipalities.

The population registries are supplemented with variables originating from sample surveys, because not all variables that are necessary according to the EU regulations can be found in the population registries. The use of data originating from registries is preferred over data originating from sample surveys, because registries generally have a better quality than surveys and because registries often contain the population instead of only a sample. In the 2001 and 2011 Dutch censuses, only two variables were not taken from registries: *Occupation* and *Educational Attainment* (Schulte Nordholt et al., 2014).

To obtain the required cross-tables for the 2011 Dutch census, a procedure was used where all data sources are matched on the unit level. Then, a micro integration process was carried out. Micro-integration entails checking and completing the data. If data are missing in one source another source can be used. Also, inconsistencies between sources are removed, by using so-called edit rules: formal rules that make clear what happens in case of inconsistencies, e.g. which source is used (Bakker, 2010; De Waal et al., 2011). It is widely assumed that using micro integration increases the quality of the data, as it combines the information from all available data sources.

After micro-integration, two combined data sources are obtained. The first is based on a combination of registries and the second is based on a combination of sample surveys. The required tables that can be produced from the combined registries are simply produced by counting from this combined data source. Census tables that need to be compiled from sample survey information are produced by weighting.

In the specific case of the Dutch population census, an important survey used is the Labor Force Survey (LFS). The LFS respondents are weighted to represent the part of the Dutch population that is not covered by the LFS sample surveys. These weights are calibrated, such that the marginal totals of the weighted data comply with the marginal totals that are known from registers. When multiple tables are estimated from sample survey data, there is a risk that common marginal totals in multiple tables are not estimated the same. This problem is solved by using the so-called repeated weighting method, as introduced by Statistics Netherlands (Houbiers, 2004).

The current procedure for obtaining the required cross-tables, which combines edit restrictions and repeated weighting, knows a number of drawbacks. First, the assumption is often made that the population registries are free of error. When a variable is measured both in the population registry and in a sample survey and the scores on these variables contradict each other, the registry score usually overrides the survey score because of this assumption. However, in practice register variables usually do contain some errors. As a consequence, an unnecessary use of edit rules can be induced. Second, the uncertainty due to the missing and conflicting values in the linked dataset used to estimate the cross-tables is not directly incorporated in the estimate of the variance. Third, the data processing procedure currently used contains a specific sequence of steps, where decisions made at specific steps are influenced by decisions made at previous steps.

In Chapter 2, a method is described that circumvents the above mentioned issues by combining Multiple Imputation and Latent Class analysis (MILC). The main goal of the MILC method is to correct for misclassification by using measures of the same variable originating from different sources (population registry and sample survey) that are linked on a unit level. These variables are used as indicators of a Latent Class (LC) model. The MILC method simultaneously corrects for misclassification detected by evaluating logically impossible combinations of scores, because edit restrictions can be incorporated into the LC model. After the LC model has been estimated, multiply imputed versions of the variable are created, that are corrected for the estimated misclassification. The differences between

the imputations reflect the uncertainty due to missing and conflicting values which can be incorporated in the estimate of the total variance. In a simulation study on the performance of this method, it was concluded that its performance is strongly related to the entropy R^2 value of the LC model, which indicates how well the LC model can predict class membership based on the observed variables, or on how well classes are separated. It was also concluded that the population registries and sample surveys used at Statistics Netherlands are generally of sufficient quality for the MILC method to obtain reliable estimates.

After the method was introduced, multiple studies have extended the method to meet requirements for specific applications. In Chapter 3 the method was extended in such a way that it could simultaneously impute values missing by design by using a quasi-latent variable. In addition, they investigated the performance of the method when two combined sources are both incomplete and follow a different missingness mechanism. Furthermore, Chapter 4 investigated how the method can be extended to longitudinal situations and how it can use a combined dataset consisting of both survey and registry data.

Although these previous studies investigated a number of relevant issues regarding the production of official statistics using combined datasets, the specific situation of using combined survey-register data for population censuses entails a number of typical characteristics for which it remains unclear how the MILC method should handle them. A first issue is the simultaneous imputation of multiple latent variables. As population registries can contain misclassification, it is worthwhile to correct for the misclassification if possible. If this is the case for multiple variables, correction should be performed simultaneously. Second, it has been shown that edit restrictions can be incorporated in an LC model to prevent the occurrence of logically impossible combinations of scores (Chapter 2). However, it is unclear how the MILC method performs if a large number of edit restrictions is incorporated, affecting multiple cells in a population census table. Third, when using a combined population registry - sample survey dataset, Statistics Netherlands generally assumes that the population registry is complete and contains all observations of a finite population. It is unclear if the MILC method can be applied directly while making such an assumption, or that adaptations to the method should be made.

Before using the MILC method to construct population census tables, it should be investigated if and how the MILC method is able to handle the above mentioned issues. Therefore, we perform a simulation study where an existing population census cross-table is assumed to be a complete and finite population. From here, misclassification and missing values are

generated to evaluate the performance of the MILC method. This enables us to evaluate the performance of the MILC method when the observed registry is assumed to be a complete and finite population, when multiple latent variables are imputed simultaneously and when edit restrictions are applied to multiple cells within the population census table simultaneously.

In Section 7.2, a description of the MILC method is given, tailored to handle the specific situation discussed. In Section 7.3, a description of the performed simulation study is given. In Section 7.4, simulation results are shown and Section 7.5 provides a discussion.

7.2 Methodology

When applying the MILC method, the starting point is the unit-linked combined dataset. In order to account for parameter uncertainty, a non-parametric bootstrap procedure is applied on this dataset first. This involves creating m bootstrap samples by drawing observations from the observed dataset with replacement. Subsequently, for each bootstrap sample, the LC model of interest is estimated (step *two*) and m imputations are created using the m sets of parameter values obtained from the m latent class models (step *three*). If imputations would be created based on the maximum-likelihood estimates obtained using the observed data, sampling uncertainty regarding the estimated parameters of the latent class model would be ignored.

7.2.1 Step 1: Creating bootstrap samples

As we consider a situation of finite population in this study, it would be incorrect to generate bootstrap samples in the usual way. In that case, it would be assumed that the parameters of interest are those of a super-population model. For a population census, the parameters of interest are those of the finite population itself. Therefore, theoretically there is no estimation uncertainty regarding population parameters that are completely observed. As it would be theoretically incorrect to sample with replacement from a complete and finite population, this is handled in a straightforward way by generating m bootstrap samples from the subset of persons observed in both the survey and the registry. The persons who are not observed in the survey dataset are considered as persons with missing values in the combined dataset. The estimation of the LC models (step *two*) is applied only on the subset of complete observations and these m sets of parameters are then used to generate m sets of posterior membership probabilities for the original datasets to be used for imputation.

7.2.2 Step 2: Estimating the latent class model

The *second* step performed is the estimation of the LC model. As described in the previous section, the LC model is typically estimated m times using the m bootstrapped datasets. In the situation under evaluation in this paper, the LC model is estimated m times on m subsets of complete observations coming from the m bootstrap samples. An extensive discussion of the model and the assumptions made when using the model to correct for measurement error can be found in Chapter 2. Multiple latent variables can be estimated simultaneously in one model, which yields the following model structure for the joint probability of the response variables given covariate values, denoted by $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})$. The number of latent variables is denoted as v and K_i is the number of classes of latent variable X_i, where $(i = 1, ..., v)$. Furthermore, $\mathbf{Y}$ are the indicator variables, L_i is the number of indicator variables for X_i and $\mathbf{Q}$ are the covariate variables:

$$\begin{aligned}\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q}) = \sum_{x_1=1}^{K_1} ... \sum_{x_v=1}^{K_v} \mathrm{P}(X_1 = x_1, ..., X_v = x_v \mid \mathbf{Q} = \mathbf{q}) \\ \prod_{l_1=1}^{L_1} \mathrm{P}(Y_{l_1,1} = y_{l_1,1} \mid X_1 = x_1) \\ ... \\ \prod_{l_v=1}^{L_v} \mathrm{P}(Y_{l_v,v} = y_{l_v,v} \mid X_v = x_v).\end{aligned} \tag{7.1}$$

Here, local independence is assumed as well as independence of covariates.

Constrained parameter estimation is used when certain cells within $\mathrm{P}(X_1 = x_1, ..., X_v = x_v \mid \mathbf{Q} = \mathbf{q})$ are restricted. This can be used to specify that certain combinations of scores between covariates and latent variables are logically impossible, or when a 'quasi-latent' variable is used to create imputations for missing values in a variable (Vermunt & Magidson, 2013b).

7.2.3 Multiple imputation

To be able to create multiple imputations, joint posterior membership probabilities are calculated for every person in the original dataset. They represent the probability that a unit is part of a combination of latent classes from the different latent variables, given its combination of scores on the indicators and covariates used in the LC model. These probabilities can be used to create multiple imputations of the latent variables which contain their *true* scores.

The joint posterior membership probabilities can be calculated by applying Bayes' rule to the conditional response probabilities obtained from the m LC models:

$$\mathrm{P}(X_1 = x_1, ..., X_v = x_v \mid \mathbf{Y} = \mathbf{y}, \mathbf{Q} = \mathbf{q}) = \frac{\mathrm{P}(X_1 = x_1, ..., X_v = x_v, \mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})}{\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})},$$

where

$$\begin{aligned}\mathrm{P}(X_1 = x_1, ..., X_v = x_v, \mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q}) = \mathrm{P}(X_1 = x_1, ..., X_v = x_v | \mathbf{Q} = \mathbf{q}) \\ \prod_{l_1=1}^{L_1} \mathrm{P}(Y_{l_1,1} = y_{l_1,1} \mid X_1 = x_1) \\ ... \\ \prod_{l_v=1}^{L_v} \mathrm{P}(Y_{l_v,v} = y_{l_v,v} \mid X_v = x_v)\end{aligned}$$

and $\mathrm{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})$ is defined in Equation 7.1. For one profile (so one set of scores on all indicator and covariate variables), the joint posterior membership probabilities sum up to one.

We have to estimate the m LC models on the original and complete dataset. As the m LC models are estimated using the complete observations of the m bootstrap samples of the original dataset, the sets of parameters obtained from these m LC models can be used as starting values of a new LC model using the original complete dataset (including persons containing missing values) where iterations of the EM-algorithm are suppressed to zero. In this way, the LC model is calculated directly using the parameters obtained from the bootstrap sample in one E-step of the algorithm.

7.2.4 Pooling

The next step is to obtain estimates of interest for every imputation, and to pool them using Rubin's Rules (Rubin, 1987, p.76). For this research, the main interest is producing a frequency table. Therefore, the frequency table of interest is obtained for the m imputations and they are pooled, which means taking the average over the imputations for every cell in the frequency table:

$$\hat{\theta}_j = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_{ij},$$

where j refers to a specific cell in the frequency table.

Next, an estimate of the uncertainty around these frequencies is of interest. As the population

is finite and observed and imputations for the complete population are generated, sampling variance should not be incorporated here (Vink & van Buuren, 2014). Therefore, the only cause of variance is uncertainty due to missing and conflicting values in the indicator variables. This should be reflected by the differences between the m imputations. Therefore, the within variance component of Rubins' total variance formula for multiple imputation (Rubin, 1987, p.76) can be mitigated and

$$\text{VAR}_{\text{total}_j} = \overline{\text{VAR}}_{\text{within}_j} + \text{VAR}_{\text{between}_j} + \frac{\text{VAR}_{\text{between}_j}}{m}$$

can be reduced to

$$\text{VAR}_{\text{total}_j} = \text{VAR}_{\text{between}_j} + \frac{\text{VAR}_{\text{between}_j}}{m}.$$

Here, $\text{VAR}_{\text{between}_j}$ can be estimated as

$$\text{VAR}_{\text{between}_j} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_{ij} - \hat{\theta}_j)(\hat{\theta}_{ij} - \hat{\theta}_j)'. \tag{7.2}$$

7.3 Simulation study

The goal of this paper is to investigate whether the MILC method is able to produce a population census table using a contaminated administrative data source and a survey sample containing misclassification. In this section, we describe the simulation study that is performed to evaluate this situation.

7.3.1 The census table used as starting point

The starting point of this simulation study is an existing census table, which can be downloaded from Census Hub (Census Hub, 2017). This table comprises $2,691,477$ persons who where living in the region 'Noord-Holland' in the Netherlands in 2011. This census table is a cross-table between the following six variables:

1. *Age* in 21 categories: under 5 years; 5 to 9 years; 10 to 14 years; 15 to 19 years; 20 to 24 years; 25 to 29 years; 30 to 34 years; 35 to 39 years; 40 to 44 years; 45 to 49 years; 50 to 54 years; 55 to 59 years; 60 to 64 years; 65 to 69 years; 70 to 74 years; 75 to 79 years; 80 to 84 years; 85 to 89 years; 90 to 94 years; 95 to 99 years; 100 years and over.

2. *Marital status* in eight categories: never married; married; widowed; divorced; registered partnership; widow of registered partner; divorced from registered partner;

not stated.

3. *Gender* in two categories: male; female.
4. Place of birth in five categories: the Netherlands; a country within the European Union; a country outside the European Union; other; not stated.
5. *Type of family nucleus* in which a person lives in five categories: partners; lone parents; sons/daughters; not stated; not applicable.
6. *Country of citizenship* in five categories: Dutch citizen; citizen of a country within the European Union; citizen of a country outside the European Union; stateless; not stated.

Thus, the census table consists of $42,000$ cells.

7.3.2 Simulation setup

The goal of this simulation study is to replicate the frequencies of the $42,000$ cells in the cross-table using multiple indicators contaminated with misclassification and missing values. The misclassification is induced first.

We generate two indicator variables for three different latent variables, all containing 5% random misclassification, which can be considered a high amount, especially for Dutch population registries. The indicator variables are generated for the variables *Gender*, *Type of family nucleus* and *Country of citizenship*. Misclassification is generated in such a way that first, 5% of the cases are randomly selected. Second, their original score is identified and third, a different score is assigned by sampling from the observed frequency distribution of the other categories.

For the first indicator of each variable, misclassification is generated only once, as these indicator variables represent register variables for the complete and finite population, there should not be any variability in misclassification between replications in the simulation study for these variables. For the second indicator of each variable, misclassification is newly generated for every replication in the simulation study, followed by generating missing values using either a Missing Completely At Random (MCAR) or Missing At Random (MAR) missingness mechanism with approximately 90% missingness for both situations. These second indicators represent survey variables for a sample of the population.

Missingness is generated in such a way that it mimics a situation that 10% of the population is included in the survey. Missingness is generated under MCAR and MAR. Under MCAR, the

probability of being missing (i.e. not being included in the survey) is 90% and equal for every person in the population. Under MAR, the probability of being missing depends on a persons' age and decreases as a person gets older. More specifically, the probability of being missing is lowest for persons in the age category '100 years and older', and is 80%. This percentage gradually increases with the highest being 94% for the persons in the age category 'under five years'.

7.3.3 Applying the MILC method

As discussed in Section 7.2, m bootstrap samples are generated from the combined dataset, and the LC model is estimated only on the complete set of observations of each bootstrap sample. Results are obtained using $m = 5$, $m = 10$ and $m = 20$.

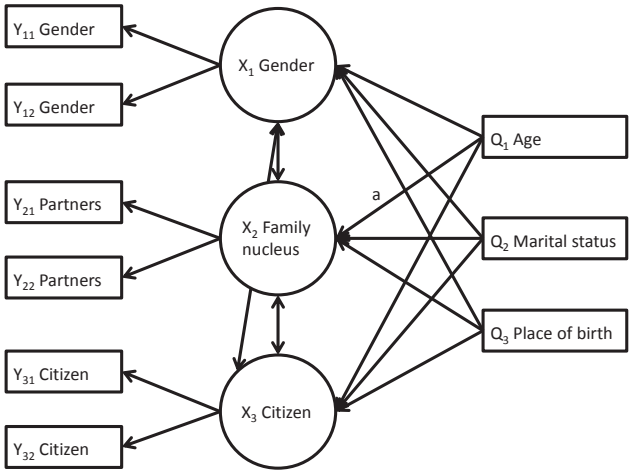

Figure 7.1: Graphical overview of the LC model specified. Note that edit restrictions are applied between the variables *Type of family nucleus* and *Age* (denoted in the model by 'a').

In Figure 7.1, the graphical overview of the latent class model can be found. Here, it can be seen that the latent variables X_1 *Gender*, X_2 *Family nucleus* and X_3 *Citizen* are all measured by two indicators. The restriction on the relationship between Q_1 *Age* and X_2 *Family nucleus* is denoted by 'a' in Figure 7.1. Here, we restricted that if someone is of *Age* category 'under 5 years', '5 to 9 years' or '10 to 14 years', it is impossible to be assigned to the latent classes 'partners' or 'lone parents' for the latent variable *Family nucleus*.

To specify the LC model for response pattern $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{Q} = \mathbf{q})$ we can fill in at Equation 7.1 that $v = 3$, $K_1 = 2$, $K_2 = 4$, $K_3 = 4$, $L_1 = 2$, $L_2 = 2$ and $L_3 = 2$. Note that X_2 here only has four latent classes, while the variable *Family Nucleus* in the population census table has five categories. Therefore, it would have made sense for X_2 to also have five latent classes. However, there were no observations for the category 'not applicable', so therefore a latent class is not included for this category. The same holds for the category 'stateless' of X_3.

Next, multiple imputations can be created and estimates of interest can be pooled as described in Section 7.2.3 and Section 7.2.4. As the cells of the frequency-tables of interest can become very small, a log-transformation is used to ensure appropriate confidence intervals around these small cells. Therefore, $\text{VAR}_{\text{between}_j}$ is not estimated as the variance of $\hat{\theta}_{ij}$, as in Equation 7.2, but as the variance of $\log(\hat{\theta}_{ij})$, where $\hat{\theta}_{ij}$ refers to the number of units in cell j in imputation i.

7.3.4 Evaluation

To evaluate the performance of the MILC method when trying to construct the census table initially used to create the misclassified data, it is useful to make comparisons to results obtained when the variable observed in the registry is used directly to create cross-tables. We refer to these results as obtained using $Y_{\mathbf{v},1}$. Furthermore, it would be difficult to draw general conclusions from results obtained by only evaluating every single of the $42,000$ cells of the complete census table. Therefore, some specific characteristics of this table are investigated separately. First, it is investigated whether the method is able to reconstruct the univariate marginal cell frequencies of the latent variables specified. Second, it is investigated if the method is able to reconstruct the joint distribution of the three latent variables. Third, it is investigated if the method correctly incorporates edit restrictions. At last, some features of the complete census table are investigated.

First, the cell-proportions of the previously discussed cross-tables are investiaged in terms of bias, by evaluating the average absolute bias and the root mean squared error (RMSE) over the 500 replications in the simulation study. Second, results are evaluated in terms of variance. Here, it is of interest to evaluate whether MILC correctly reflects uncertainty due to missing and conflicting values in between imputation variance for both univariate and multivariate cross-tables. Therefore, it is investigated if the average of the estimated standard errors is approximately equal to the standard deviation over the 500 estimates obtained from the 500 simulation replications by evaluating its ratio.

To summarize, the specific conditions evaluated in this simulation study are denoted as $Y_{v,1}$, MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20, MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.

7.4 Simulation results

First, cell-proportions of univariate and multivariate cross-tables are evaluated in terms of bias and root mean squared error (RMSE) over the 500 simulation replications. Second, these cell-proportions are evaluated in terms of variance by investigating the average of the estimated standard error divided by the standard deviation over the 500 estimates obtained from the 500 simulation replications.

7.4.1 Results in terms of bias

7.4.1.1 Univariate marginal frequencies of imputed variables

Table 7.1: Results in terms of bias and root mean squared error for the two categories of the imputed latent variable *Gender* ('F.' is 'Female' and 'M.' is 'Male').

Gender								
				MCAR			MAR	
	Frequency	$Y_{1,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias				
F.	1,367,167	-2,126	3,386	3,308	3,325	3,231	3,153	3,109
M.	1,324,310	2,126	-3,386	-3,308	-3,325	-3,231	-3,153	-3,109
				RMSE				
F.	1,367,167	2,154	6,008	5,888	5,760	5,914	5,637	5,512
M.	1,324,310	2,154	6,008	5,888	5,760	5,914	5,637	5,512

In Table 7.1, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable *Gender* in terms of bias and RMSE. Results from all simulation conditions are shown. Here, it can be seen that a smaller amount of bias is obtained if $Y_{1,1}$ is used, compared to results obtained using MILC under all conditions. In addition, it can be seen that the RMSE is also smaller if $Y_{1,1}$ is used instead of the MILC method. Furthermore, it can be seen that both bias and RMSE slightly decrease as m increases, and that the quality of the results appears to be unrelated to the missingness mechanism.

In Table 7.2, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable *Type of family nucleus* in terms of bias and RMSE. Here, the results are very different from the results we found for *Gender*, the bias obtained for $Y_{2,1}$ much higher compared to the bias obtained using MILC under all conditions and the same holds

Table 7.2: Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable *Type of family nucleus* ('N.A.' means 'Not applicable'). Note that the category 'Not stated' is mitigated as it contained zero observations.

Type of family nucleus								
				MCAR			MAR	
	Frequency	$Y_{2,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias				
Lone parents	97,360	2,670	185	182	176	224	226	220
N.A.	604,032	8,985	-957	-975	-989	-1,601	-1,612	-1,611
Partners	1,272,339	-19,686	401	411	427	932	935	932
Sons/daughters	717,746	8,030	371	381	386	446	451	459
				RMSE				
Lone parents	97,360	2,672	425	408	395	426	421	414
N.A.	604,032	8,989	1,337	1,318	1,312	1,837	1,833	1,818
Partners	1,272,339	19,688	954	914	904	1,256	1,235	1,218
Sons/daughters	717,746	8,034	630	624	617	715	692	688

for RMSE. In addition, whether the results for the MILC method depend on the missingness mechanism differ per category. In terms of bias and RMSE, this is the case for the categories 'N.A.' and 'Partners'.

Table 7.3: Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable *Citizen* ('N.S.' means 'Not stated'). Note that the category 'Stateless' is mitigated as it contained zero observations.

Citizen								
				MCAR			MAR	
	Frequency	$Y_{3,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias				
EU	79,212	51,365	-5	-7	-12	-199	-211	-216
NL	2,511,214	-116,899	-555	-546	-545	117	124	107
not EU	89,592	58,085	512	502	507	62	69	89
Not stated	11,459	7,448	49	51	49	21	18	20
				RMSE				
EU	79,212	51,365	410	398	388	488	486	475
NL	2,511,214	116,899	925	894	883	767	756	720
not EU	89,592	58,086	800	770	767	618	611	590
Not stated	11,459	7,449	201	197	190	204	205	198

In Table 7.3, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable *Citizen* in terms of bias and RMSE. Here, the results are comparable to the results we found for *Type of family nucleus*, as the bias obtained when only $Y_{3,1}$ is used is again much higher compared to the bias obtained using MILC method and the

same holds for RMSE. As was also the case for *Type of family nucleus*, whether the results for the MILC method depend on the missingness mechanism differ per category, although this is more the case for bias here, and not so much in terms of RMSE.

In Chapter 2 it was concluded that the quality of the output when MILC is applied is related to how well the latent class model is able to make classifications based on the observed data, which is summarized in the entropy R^2. The entropy R^2 values for *Gender*, *Type of family nucleus* and *Citizen* are approximately 0.7352, 0.9191, and 0.8571 respectively under MCAR. So this corresponds to the quality of the results for the latent variables in terms of bias and RMSE. An additional explanation for *Gender* is that the two categories are of comparable size and the amount of misclassification in both categories is approximately equal and behaves symmetrical in our simulation study. This causes that the marginal distribution of $Y_{1,1}$ is very similar to the marginal distribution of X_1 and not so much affected by misclassification.

7.4.1.2 Joint frequencies of imputed variables

In Table 7.4 and Table 7.5, the simulation results can be found that cover the joint marginal frequencies of the three imputed latent variables in terms of bias and RMSE. Again, it can be seen here that if only $Y_{v,1}$ is used, severe bias is present in all cells of the joint frequency table. The results obtained when the MILC method is applied show much lower amounts of bias and RMSE. Here, the differences between different numbers for m or different missingness mechanism are much smaller compared to the differences between MILC and $Y_{v,1}$.

7.4.1.3 Restricted cells

In Table 7.6, the simulation results can be found for the six cells that are restricted in the marginal cross-table between *Age* and *Type of family nucleus*. Under 'Frequency', it can be seen that these six cells should all contain zero observations. A combination of these scores is logically impossible. Furthermore, it can be seen that due to misclassification in $Y_{2,1}$, observations containing these combinations of scores are present when $Y_{2,1}$ is used to estimate this cross-table directly. In addition, it can be seen that if the MILC method is applied, such impossible combinations of scores will never be present, regardless of the missingness mechanism or the number of imputations. Furthermore, as the cells in this marginal table contain zero observations, all cells of more detailed tables covering these logically impossible combinations of scores automatically also contain zero observations.

Table 7.4: Results in terms of bias for the 32 observed categories of the joint distribution of the three imputed latent variables *Gender*, *Type of family nucleus* and *Citizen*. For *Gender* ('G.'), 'F.' means Female and 'M.' means Male. For *Type of family nucleus* ('F.N.'), 'L.P.' means Lone Parents, 'N.A.' means Not applicable, 'P.' means Partners and 'S.D.' means Sons/Daughters. 'N.S.' means 'Not stated'. Note that the categories 'Stateless' for *Citizen* and 'Not Stated' for *Type of family nucleus* are mitigated as they contained zero observations.

Gender × *Type of family nucleus* × *Citizen*										
				MCAR				MAR		
G.	F.N.	C.	freq.	$Y_{\mathbf{v},1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias						
F.	L.P.	EU	2,091	1,434	8	7	7	1	0	0
F.	L.P.	NL	76,131	-6,620	652	650	646	240	241	234
F.	L.P.	not EU	3,120	1,513	33	32	32	39	39	38
F.	L.P.	N.S.	646	154	-5	-5	-6	-13	-13	-13
F.	N.A.	EU	12,436	5,971	433	432	432	431	427	427
F.	N.A.	NL	293,960	-11,998	-595	-618	-623	905	891	880
F.	N.A.	not EU	9,509	7,317	1,032	1,031	1,032	1,069	1,069	1,071
F.	N.A.	N.S.	1,221	982	182	182	182	198	197	197
F.	P.	EU	20,443	11,185	237	236	235	24	19	21
F.	P.	NL	584,547	-34,001	294	262	279	-564	-599	-624
F.	P.	not EU	26,877	12,022	404	402	401	254	255	258
F.	P.	N.S.	1,292	1,837	-19	-18	-18	-23	-24	-24
F.	S.D.	EU	4,368	7,541	-778	-779	-780	-851	-853	-854
F.	S.D.	NL	321,364	-8,738	2,483	2,471	2,479	2,620	2,601	2,588
F.	S.D.	not EU	7,680	8,303	-764	-768	-766	-876	-874	-869
F.	S.D.	N.S.	1,482	971	-209	-208	-208	-223	-223	-222
M.	L.P.	EU	389	591	-10	-11	-11	9	9	9
M.	L.P.	NL	14,536	4,791	-553	-552	-554	-134	-131	-130
M.	L.P.	not EU	372	707	35	35	35	53	53	53
M.	L.P.	N.S.	75	100	27	27	27	28	29	29
M.	N.A.	EU	16,308	4,444	-306	-304	-305	-349	-349	-350
M.	N.A.	NL	253,493	-3,733	-714	-708	-717	-2,730	-2,722	-2,713
M.	N.A.	not EU	13,636	5,548	-904	-903	-902	-1,023	-1,023	-1,020
M.	N.A.	N.S.	3,469	455	-85	-86	-87	-102	-103	-104
M.	P.	EU	18,444	11,881	793	796	794	905	906	906
M.	P.	NL	599,278	-38,164	-3,170	-3,128	-3,127	-1,528	-1,490	-1,474
M.	P.	not EU	19,776	13,709	1,794	1,793	1,793	1,785	1,790	1,791
M.	P.	N.S.	1,682	1,846	69	69	69	78	78	79
M.	S.D.	EU	4,733	8,319	-382	-382	-384	-370	-371	-374
M.	S.D.	NL	367,905	-18,435	1,049	1,076	1,072	1,308	1,333	1,346
M.	S.D.	not EU	8,622	8,966	-1,118	-1,120	-1,117	-1,240	-1,239	-1,233
M.	S.D.	N.S.	1,592	1,103	90	90	91	77	77	78

7.4.1.4 The complete population frequency table

Figure 7.2 and Figure 7.3 show results in terms of bias and root mean squared error (RMSE) when the complete census table, so the cross-table between the six variables, is estimated. As these are $42,000$ cells in total, it is not feasible to evaluate them individually. Figure 7.2 and Figure 7.3 give an overview of how size of the cell frequency is related to the quality of

Table 7.5: Results in terms of root mean squared error for the 32 observed categories of the joint distribution of the three imputed latent variables '*Gender*, *Type of family nucleus* and *Citizen*. For *Gender* ('G.'), 'F.' means Female and 'M.' means Male. For *Type of family nucleus* ('F.N.'), 'L.P.' means Lone Parents, 'N.A.' means Not applicable, 'P.' means Partners and 'S.D.' means Sons/Daughters. 'N.S.' means 'Not stated'. Note that the categories 'Stateless' for *Citizen* and 'Not Stated' for *Type of family nucleus* are mitigated as they contained zero observations.

Gender × *Type of family nucleus* × *Citizen*										
					MCAR			MAR		
G.	F.N.	C.	freq.	$Y_{\mathbf{v},1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				RMSE						
F.	L.P.	EU	2,091	1,434	45	42	41	45	42	40
F.	L.P.	NL	76,131	6,621	742	734	724	418	408	394
F.	L.P.	not EU	3,120	1,514	67	64	64	71	68	66
F.	L.P.	N.S.	646	155	22	21	20	26	25	24
F.	N.A.	EU	12,436	5,972	449	446	445	447	442	440
F.	N.A.	NL	293,960	12,001	1,260	1,245	1,222	1,433	1,374	1,348
F.	N.A.	not EU	9,509	7,317	1,038	1,037	1,037	1,075	1,075	1,076
F.	N.A.	N.S.	1,221	983	185	185	185	202	201	201
F.	P.	EU	20,443	11,186	291	285	282	173	163	157
F.	P.	NL	584,547	34,003	2,332	2,285	2,204	2,364	2,248	2,197
F.	P.	not EU	26,877	12,023	456	450	447	330	327	327
F.	P.	N.S.	1,292	1,838	46	44	43	48	48	47
F.	S.D.	EU	4,368	7,541	787	787	787	860	862	863
F.	S.D.	NL	321,364	8,742	2,820	2,796	2,781	2,959	2,903	2,879
F.	S.D.	not EU	7,680	8,304	779	782	780	892	889	883
F.	S.D.	N.S.	1,482	972	216	214	214	230	230	229
M.	L.P.	EU	389	592	18	17	17	17	17	16
M.	L.P.	NL	14,536	4,792	605	600	600	271	260	257
M.	L.P.	not EU	372	707	38	38	37	55	55	55
M.	L.P.	N.S.	75	101	27	27	27	29	29	29
M.	N.A.	EU	16,308	4,445	331	328	327	373	371	370
M.	N.A.	NL	253,493	3,742	1,390	1,349	1,314	2,959	2,931	2,911
M.	N.A.	not EU	13,636	5,549	913	912	911	1,033	1,031	1,028
M.	N.A.	N.S.	3,469	456	107	105	104	121	121	120
M.	P.	EU	18,444	11,881	808	810	807	919	919	917
M.	P.	NL	599,278	38,165	3,898	3,837	3,794	2,755	2,617	2,568
M.	P.	not EU	19,776	13,709	1,804	1,803	1,803	1,797	1,800	1,800
M.	P.	N.S.	1,682	1,846	88	87	85	98	95	95
M.	S.D.	EU	4,733	8,319	403	403	403	401	401	402
M.	S.D.	NL	367,905	18,437	1,728	1,723	1,687	1,905	1,872	1,854
M.	S.D.	not EU	8,622	8,967	1,129	1,130	1,127	1,252	1,250	1,244
M.	S.D.	N.S.	1,592	1,104	109	108	107	103	102	101

the results. Here it can be seen that if $Y_{v,1}$ are used, results in terms of bias and RMSE are related directly to cell frequency. More specifically, the relationship between cell frequency and absolute bias is approximately linear where the amount of bias is approximately 10% of the cell frequency.

Table 7.6: Results in terms of bias and root mean squared error for the six restricted categories from cross-table between 'Type of family nucleus' and the covariate *Age*. For *Type of family nucleus*, 'L.P.' means Lone Parents, and 'P.' means Partners.

Type of family nucleus									
				MCAR			MAR		
		Freq.	$Y_{2,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias					
L.P.	under 5 y.	0	377	0	0	0	0	0	0
L.P.	5 to 9 y.	0	386	0	0	0	0	0	0
L.P.	10 to 14 y.	0	376	0	0	0	0	0	0
P.	under 5 y.	0	4,934	0	0	0	0	0	0
P.	5 to 9 y.	0	5,041	0	0	0	0	0	0
P.	10 to 14 y.	0	4,937	0	0	0	0	0	0
				RMSE					
L.P.	under 5 y.	0	377	0	0	0	0	0	0
L.P.	5 to 9 y.	0	386	0	0	0	0	0	0
L.P.	10 to 14 y.	0	377	0	0	0	0	0	0
P.	under 5 y.	0	4,934	0	0	0	0	0	0
P.	5 to 9 y.	0	5,041	0	0	0	0	0	0
P.	10 to 14 y.	0	4,937	0	0	0	0	0	0

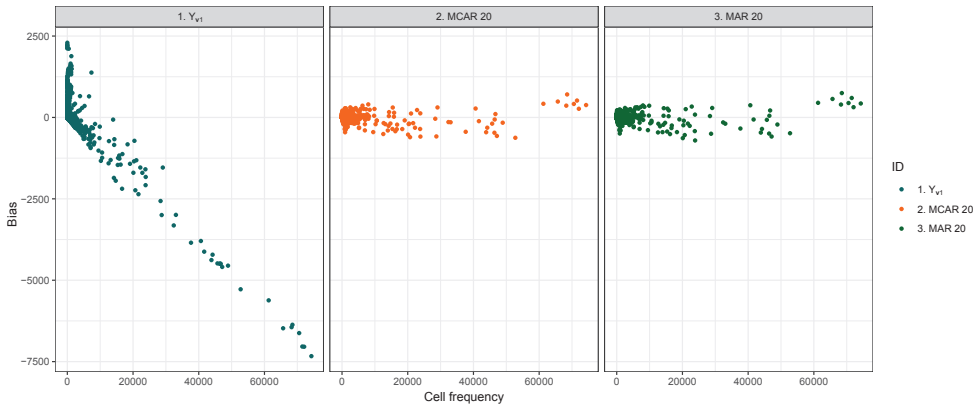

Figure 7.2: Results in terms of bias when the complete cross-table between the latent variables *Gender*, *Type of family nucleus* and *Citizen* and the three covariates *Age*, *Marital status* and *Place of birth* is estimated. The X-axis represents cell frequency and the Y-axis represents the bias. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.

7.4.2 Results in terms of variance

7.4.2.1 Univariate marginal frequencies of imputed variables

In Table 7.7, the simulation results can be found that cover the univariate marginal frequencies *Gender* in terms of se/sd. As this ratio confirms that the average standard error estimated at

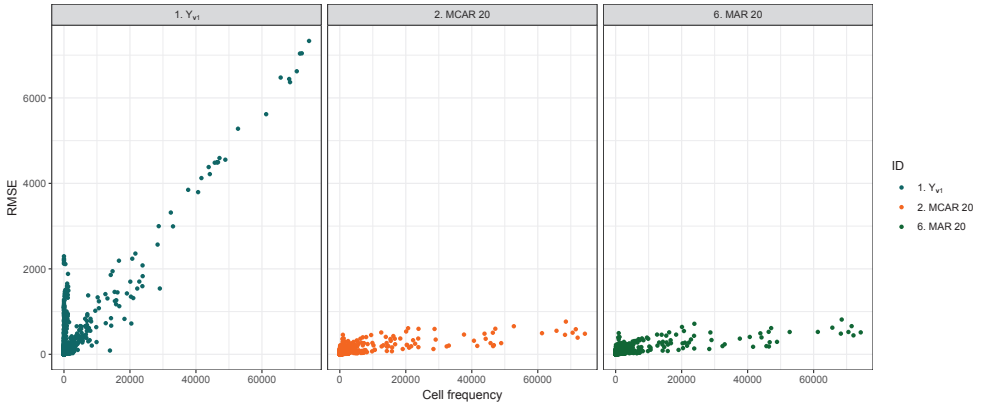

Figure 7.3: Results in terms of root mean squared error when the complete cross-table between the latent variables *Gender*, *Type of family nucleus* and *Citizen* and the three covariates *Age*, *Marital status* and *Place of birth* is estimated. The X-axis represents cell frequency and the Y-axis represents the bias. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.

Table 7.7: Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the two categories of the imputed latent variable *Gender* ('F.' is 'Female' and 'M.' is 'Male').

Gender								
				MCAR			MAR	
	Freq.	$Y_{v,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				se/sd				
F.	1,367,167	-	1.0540	1.0317	1.0363	1.0030	1.0235	1.0237
M.	1,324,310	-	1.0546	1.0317	1.0363	1.0034	1.0236	1.0236

each replication in the simulation correctly describes the uncertainty (standard deviation) that is found over the estimates, it should be close to one. In addition, as a completely observed and finite population is assumed, variance is not estimated when $Y_{v,1}$ is used. The results obtained using MILC are generally close to one and comparable to the results in terms of bias as only minor differences can be found between different values for m or between the different missingness mechanisms.

In Table 7.8 and Table 7.9, the simulation results can be found that cover the univariate marginal frequencies for *Type of family nucleus* and *Citizen* respectively in terms of se/sd. The results found here have a very comparable structure compared to the results we found for *Gender*.

Table 7.8: Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the four observed categories of the imputed latent variable *Type of family nucleus*. For *Type of family nucleus* ('F.N.'), 'L.P.' means Lone Parents, 'N.A.' means Not applicable, 'P.' means Partners and 'S.D.' means Sons/Daughters. Note that the category 'Not stated' is mitigated as it contained zero observations.

Type of family nucleus								
			MCAR			MAR		
	Freq.	$Y_{v,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				se/sd				
L.P.	97,360	-	1.0457	1.0510	1.0529	1.0561	1.0337	1.0336
N.A.	604,032	-	0.9706	0.9874	0.9922	0.9751	0.9829	0.9863
P.	1,272,339	-	1.0332	1.0418	1.0456	1.0052	1.0269	1.0298
S.D.	717,746	-	0.9594	0.9615	0.9606	0.9696	0.9880	0.9938

Table 7.9: Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the four observed categories of the imputed latent variable *Citizen* ('N.S.' means 'Not stated'). Note that the category 'Stateless' is mitigated as it contained zero observations.

Type of family nucleus								
			MCAR			MAR		
	Freq.	$Y_{v,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				se/sd				
EU	79,212	-	1.0417	1.0172	1.0362	1.0768	1.0539	1.0571
NL	2,511,214	-	1.0136	1.0113	1.0235	1.0925	1.0645	1.0927
not EU	89,592	-	0.9478	0.9632	0.9709	1.0282	0.9916	1.0125
N.S.	11,459	-	1.0063	1.0208	1.0238	1.1057	1.0861	1.1143

7.4.2.2 Joint frequencies of imputed variables

In Table 7.10, the simulation results can be found that cover the joint marginal frequencies of the imputed latent variables *Gender*, *Type of family nucleus* and *Citizen* in terms of absolute se/sd. The results found for these joint frequencies are very comparable to the results we found for the marginal frequencies. For cells with a relatively low frequency, it can be seen that the ratio is in general larger than one, indicating that the variance estimated for these frequencies (and thereby the differences between the imputations) incorporate more uncertainty than is actually found over different replications. Summarizing, more uncertainty than required is incorporated for the cells containing low frequencies. Results in terms for variance are not shown for the restricted cells, as the se/sd ratio cannot be evaluated here.

7.4.2.3 The complete population frequency table

In Figure 7.4, results can be found in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies estimated in the 500 replications in

Table 7.10: Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the 32 observed categories of the joint distribution of the three imputed latent variables *Gender*, *Type of family nucleus* and *Citizen*. For *Gender* ('G.'), 'F.' means Female and 'M.' means Male. For *Type of family nucleus* ('F.N.'), 'L.P.' means Lone Parents, 'N.A.' means Not applicable, 'P.' means Partners and 'S.D.' means Sons/Daughters. 'N.A.' means 'Not applicable'. Note that the categories 'Stateless' for *Citizen* and 'Not Stated' for *Type of family nucleus* are mitigated as they contained zero observations.

Gender × *Type of family nucleus* × *Citizen*										
				MCAR				MAR		
G.	F.N.	C.	Freq.	$Y_{\mathbf{v},1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
F.	L.P.	EU	2,091	-	1.1813	1.2097	1.2032	1.1495	1.1654	1.1997
F.	L.P.	NL	76,131	-	1.0371	1.0471	1.0504	1.0270	1.0252	1.0349
F.	L.P.	not EU	3,120	-	1.1659	1.1590	1.1519	1.1607	1.1634	1.1870
F.	L.P.	N.S.	646	-	1.0963	1.1004	1.1272	1.1110	1.1000	1.1054
F.	N.A.	EU	12,436	-	1.0850	1.0838	1.1172	1.0888	1.1065	1.1456
F.	N.A.	NL	293,960	-	1.0840	1.0652	1.0575	1.0158	1.0406	1.0461
F.	N.A.	not EU	9,509	-	1.1636	1.1822	1.1892	1.1574	1.1383	1.1562
F.	N.A.	N.S.	1,221	-	1.1789	1.1964	1.2097	1.1959	1.1826	1.2133
F.	P.	EU	20,443	-	1.0508	1.0537	1.0653	1.0689	1.0684	1.0925
F.	P.	NL	584,547	-	1.0313	1.0099	1.0189	1.0035	1.0253	1.0197
F.	P.	not EU	26,877	-	1.0532	1.0766	1.0720	1.0765	1.0725	1.0733
F.	P.	N.S.	1,292	-	1.1471	1.1566	1.1504	1.2157	1.1855	1.1940
F.	S.D.	EU	4,368	-	1.0135	1.0147	1.0338	1.0430	1.0518	1.0479
F.	S.D.	NL	321,364	-	1.0548	1.0379	1.0527	1.0017	1.0222	1.0221
F.	S.D.	not EU	7,680	-	0.9977	0.9966	0.9909	1.0249	1.0132	1.0416
F.	S.D.	N.S.	1,482	-	1.0344	1.0325	1.0357	1.0836	1.0688	1.0890
M.	L.P.	EU	389	-	1.3198	1.4136	1.4316	1.2941	1.3575	1.4470
M.	L.P.	NL	14,536	-	1.0784	1.0762	1.0736	1.0755	1.0690	1.0650
M.	L.P.	not EU	372	-	1.4159	1.3857	1.4511	1.4814	1.4481	1.4619
M.	L.P.	N.S.	75	-	1.4330	1.5192	1.5659	1.4598	1.5035	1.5373
M.	N.A.	EU	16,308	-	1.0990	1.0908	1.1165	1.0894	1.1022	1.1366
M.	N.A.	NL	253,493	-	1.0035	1.0100	1.0193	0.9920	1.0175	1.0238
M.	N.A.	not EU	13,636	-	1.1168	1.1100	1.1141	1.0826	1.1054	1.0952
M.	N.A.	N.S.	3,469	-	1.0241	1.0818	1.1052	1.1592	1.1478	1.1780
M.	P.	EU	18,444	-	1.1618	1.1593	1.1579	1.1473	1.1335	1.1476
M.	P.	NL	599,278	-	1.0668	1.0444	1.0487	1.0081	1.0329	1.0231
M.	P.	not EU	19,776	-	1.0932	1.0788	1.0816	1.0674	1.0612	1.0911
M.	P.	N.S.	1,682	-	1.1068	1.1411	1.1418	1.1335	1.1719	1.1770
M.	S.D.	EU	4,733	-	1.0598	1.0396	1.0548	1.0528	1.0497	1.0414
M.	S.D.	NL	367,905	-	1.0549	1.0347	1.0365	1.0098	1.0298	1.0340
M.	S.D.	not EU	8,622	-	1.0077	1.0093	1.0100	1.0413	1.0449	1.0471
M.	S.D.	N.S.	1,592	-	1.0472	1.0617	1.0699	1.0458	1.0362	1.0627

the simulation study (se/sd). Here it can be seen that the standard error estimated per cell frequency is especially too large when cell frequencies are close to zero, and become closer to

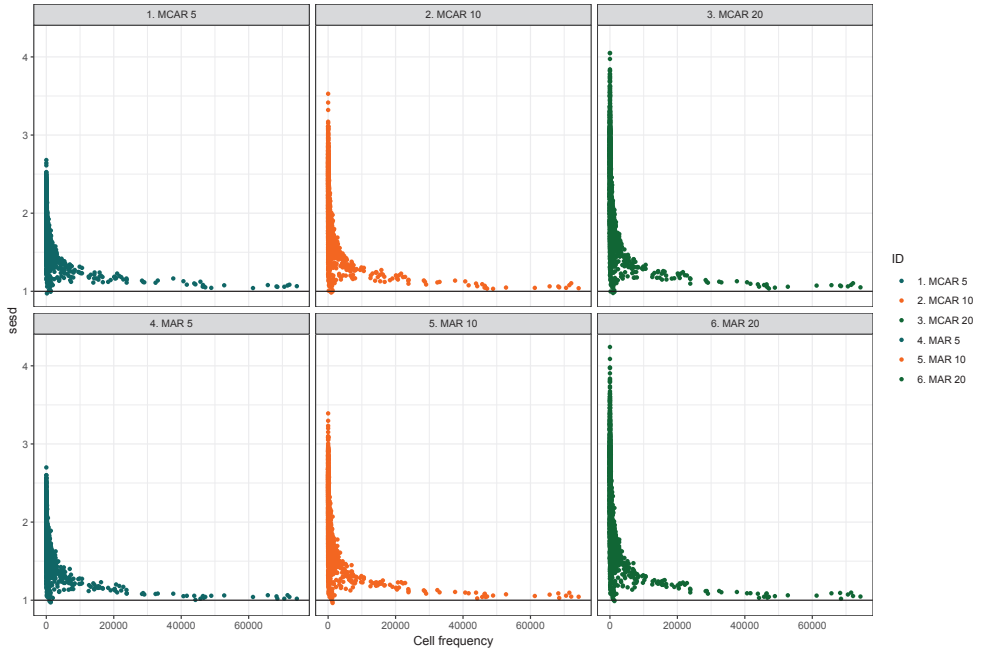

Figure 7.4: Results in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies (se/sd) when the complete cross-table between the latent variables *Gender*, *Type of family nucleus* and *Citizen* and the three covariates *Age*, *Marital status* and *Place of birth* is estimated. The X-axis represents cell frequency and the Y-axis represents the bias. Results are shown for MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20 MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.

the nominal rate of one as the cell frequencies become larger. Apparently, variability due to missing and conflicting values is overestimated by MILC for cells with a frequency close to zero. In addition, this becomes more apparent when the number of imputations increases and it is not influenced by missingness mechanism.

An additional explanation can be that the bootstrap samples are drawn with replacement and thereby assuming an infinite population. A finite population correction could correct for the fact that the sample in practice was drawn without replacement from a finite population. This would lower the estimated variances. As we currently deal with a sample of 10%, the effect of this finite population correction would be minimal. However, these bootstrap samples could explain the bias of standard errors for small cell frequencies. The bootstrap samples are drawn by resampling with replacement, while for the different replications in the simulation missingness and misclassification are generated random, however not by using a resampling mechanism.

7.5 Discussion

In this paper, the performance of the MILC method was investigated in a situation where misclassification was induced in a finite population setting. Here, an existing population census table was used as a starting point, and for three categorical variables present in this census table, two indicator variables were generated with 5% misclassification each, where one indicator also contains approximately 90% missing values. As a finite population was assumed, the estimated variance only contained a between variance component reflecting the the differences between the imputations and thereby the uncertainty caused by the misclassification and missing values in the indicator variables.

The simulation results show that the method, regardless of the number of imputations, produces results with a low bias for marginal frequency distributions, cross-tables between imputed latent variables and covariates and even for the complete six-way cross-table. Striking is the amount of bias that is induced when $Y_{\mathbf{v},1}$ is used to calculate the cross-tables evaluated in comparison to when MILC is used. It is also shown that if these indicators (i.e. $Y_{v,1}$) are used, it is likely that impossible combinations of scores are produced as well, something that can be easily circumvented by specifying edit restrictions in the LC model. This simulation study once again shows that misclassification, even if it is non-systematic, can seriously bias results. In terms of variance, it was seen that if the MILC method is applied, variance estimates are appropriate in general. However, if cell frequencies are relatively small, the variance estimates are becoming larger than necessary. This problem is more severe if the complete frequency table is evaluated, because this large table contains many cells with low frequencies.

The current set-up of this simulation study knows two major limitations. The first is caused by the large amount of cells in the cross-table. Because of this, a latent class model containing only main effects was used. It was not feasible to use a saturated model as the number of parameters would be very large, and it would be likely that not every parameter is estimable in every bootstrap sample. This would limit the use of starting values, thereby increasing the computation time for the simulation study to an unfeasible amount.

A second limitation is that we used an ad-hoc solution to handle the problem that we wanted to use bootstrapping for a subset of the dataset. More specifically, in the approach we used, we only bootstrapped the complete observations and estimated an LC model on this subset. Next, we imputed X_1, X_2 and X_3 using parameters obtained using bootstrap samples of the cases having observations in both the registry and the survey. This approach was sufficient for

our relatively simple simulation set-up, but is in theory suboptimal as it neglects information provided by the incomplete cases. A more comprehensive solution would be needed to tackle this issue.

The starting point of this simulation study was an existing population census table. A nice property here was that we could approach this as a finite and known population. Therefore, we did not have to include (within) sampling variance in our estimate of the total variance. It was insightful to evaluate cell frequencies of both univariate and multivariate cross-tables as results generally appeared to be related to cell-frequency.

Chapter 8

Summary and Discussion

8.1 Summary

National Statistical Institutes (NSIs) such as Statistics Netherlands (in Dutch: Centraal Bureau voor de Statistiek, CBS) often use large datasets to estimate population tables on many different aspects of society. A way to create these rich datasets as efficiently and cost effectively as possible is by utilizing already available population registries containing administrative data. When more information is required than already available, population registries can be supplemented with survey data. However, a major problem is that the scores of variables in both surveys and administrative data can be inconsistent and inaccurate because of various reasons, i.e. they contain misclassification.

To overcome the issue of misclassification in both kinds of sources, a method is developed in this dissertation which combines multiple imputation (MI) and Latent Class (LC) analysis (denoted as MILC). This method estimates the amount of misclassification and simultaneously imputes a new variable that is corrected for that misclassification. Furthermore, uncertainty due to misclassification is incorporated by using multiple imputations. Edit rules can be incorporated in the MILC method, which prevents impossible combinations of scores from occurring in the multiply imputed dataset.

In **Chapter 2**, MILC was introduced and a step-by-step description was given of how the method can be applied, as well as which assumptions should be met. Subsequently, the performance of the method was investigated using a simulation study. From this simulation study we were able to conclude that the quality of the results obtained after applying the MILC method depends on the entropy R^2 value, which provides a score between zero and one indicating how well you can classify data based on the observations. Therefore, it comes as no surprise that the performance of MILC is strongly related to entropy R^2. How the MILC method can be applied is also demonstrated using combined survey-registry data from CBS concerning data on home ownership.

In **Chapter 3**, the MILC method was extended to estimate the number of serious road injuries per *vehicle type* in the Netherlands on a yearly basis using two incomplete registries. Furthermore, multiple covariates were included into the model to be able to stratify the number of serious road injuries per *vehicle type* into relevant subgroups. Here, the model was extended using a 'quasi-latent variable' to be able to create multiple imputations for the covariate *region of accident*. This enabled us to estimate the number of serious road injuries per *vehicle* type per *region*, which was not possible before.

In **Chapter 4**, MILC was extended using latent Markov modeling, to be able to handle longitudinal data and correspondingly create multiple imputations for multiple time-points. A simulation study was used to investigate the performance of this extension in a situation where *true employment rates* are estimated from a combined dataset consisting of the Labor Force Survey (LFS) and population registries. Furthermore, the extended MILC method was also applied to a combined LFS-registry dataset from Italy. This simulation study showed that the MILC method can be extended to create imputations of *true* scores for multiple time-points. In addition, it showed that different strategies can be used to create the imputations, for instance by making the imputations conditional on previous imputations, or not.

In **Chapter 5** latent variable imputations were used in further analyses with covariates. These relationships can only be correctly estimated if the covariates are included in the LC model. Otherwise, point estimates will be biased. To handle situations where covariates were not included in the LC model, we extended the LC method in such a way that latent variable imputations can be updated to be conditional on the external covariates. To enable this, the MILC method was combined with two alternative approaches to 'three-step' modeling, namely the ML and BCH approach. Simulation studies were performed to investigate the performance of this extension. The main conclusion here was that the need for the correction method increases as the relationship between the imputed latent variable and the external covariate of interest is stronger. In addition, in **Chapter 6**, the BCH correction approach was placed in a framework of quadratic loss functions and linear equality and inequality constraints. In this way, inadmissible solutions with negative cell proportions can be prevented, and marginals can be fixed to specific values so that edit restrictions can be incorporated.

In **Chapter 7**, it was shown how the MILC method performs when the starting point is a population census table. In contrast to the previous chapters, the population census table was assumed to contain the finite and *true* population, which enabled us to evaluate the absolute population frequencies. Furthermore, the simultaneous imputation of multiple latent variables and application of multiple edit restrictions was evaluated. This showed that bias is generally very low when the MILC method is applied, but that in terms of variance, the MILC method can be crude, as different latent variables might require a different number of imputations. Furthermore, the chapter once again showed that simply using the error prone indicator variables and assuming that they contain the true population values, causes large amounts of bias in your results.

8.2 Discussion

Before discussing the strengths and limitations of the MILC method and our research on this method in more detail, it is important to highlight the consequences of assuming a variable to be free of misclassification and using it in further analyses. The simulation studies performed in Chapter 3 and 7 corroborate findings from earlier studies that using error prone variables without correction can result in substantive amounts of bias (Alwin, 2007, p.3-4; De Leeuw et al., 2012, p.287; Loken & Gelman, 2017). In the remainder for this section, we critically evaluate the MILC method and the research performed on the MILC method in this thesis in terms of its performance and flexibility, in terms of the assumptions made when specifying the latent class model and regarding parameter uncertainty. At last, we place MILC, as it is a misclassification correction procedure, in the Total Survey Error framework.

In this thesis, the MILC method was introduced and its performances was evaluated on a wide ranging set of conditions. The method could be adapted to the different situations in a relatively straightforward manner and its performance appeared to be strongly related to the ability of the latent class model to classify based on the observed data. While evaluating the MILC method in different settings and under different conditions, a number of issues always remained a point of discussion and they are evaluated in the following sections.

8.2.1 Performance and flexibility

The performance of the MILC method and several extensions of it, was evaluated in different simulation studies within this thesis. After the simulation study performed in Chapter 2 already, it was concluded that the performance of MILC strongly relates to the entropy R^2 value of the latent class model (indicating how well classifications can be made based on the observed data) in combination with a sufficient number of imputations. This simulation study, as well as the simulation studies performed in subsequent chapters, show that if the entropy R^2 is closer to one, the MILC method is better able to produce unbiased results with appropriate variance estimates. In addition, the simulation studies showed that the required minimum entropy R^2 value also depends on the type of output a researcher is interested in. Research on the number of imputations has been performed in Chapter 2, Chapter 3, Chapter 4 and Chapter 7, which all concluded that a low number of imputations, such as five, should generally be sufficient. Within the literature on multiple imputation for missing values, the required number of imputations is often related to the fraction of missing information (J. W. Graham et al., 2007), which is an indicator of the strength of the associations between the observations. Here, a strong association generally relates to a lower number of imputations,

as a strong association provides you with more information on the missing values. Therefore, it would be insightful to closely examine how entropy R^2 and fraction of missing information differ and/or complement each other and if an optimal number of imputations can be derived by combining the two.

In terms of flexibility of the MILC method, the latent class model can be quite easily extended to incorporate requirements from certain specific situations, as was the case in Chapter 3 and Chapter 4. In addition, covariates can also be added to the model at a later time-point as shown in Chapter 5. However, there are limitations to this flexibility. In these Chapters it was shown that the entropy R^2 values can differ for the different latent variables specified within one latent class model, resulting in the possibility that the quality of the output might differ for different imputed latent variables. In Chapter 3, the variable *region of hospital* is used as an indicator for the quasi-latent variable *region of accident*. By specifying *region of hospital* as an indicator, the opportunity to use *region of hospital* as a covariate is thereby taken away. This might be a problem if a research also has the intention to use *region of hospital* as a covariate in relation to a different latent variable used in the same model. A topic for further research is when different indicators measure the same construct with a different number of categories. Such a situation is likely to occur for example for the variable *level of education*. It would be useful to know if one variable measures *level of education* with, say, three categories, and a second variable measures *level of education* with, say, seven categories, what the appropriate number of latent classes should be in such a situation. A similar situation occurred in Chapter 4, where the 'non-motorized, other' *vehicle type* category was not measured by the police. In this situation, it was handled in a more ad-hoc manner, but it would be very useful if the latent class model itself was able to handle these different numbers of categories. A way to handle this issue within the LC framework could be to put constraints on the obvious classes so the LC model only focuses on the unspecified classes.

8.2.2 Latent class model assumptions

At all steps taken while applying the MILC method, assumptions are made. Assumptions relevant for further investigation are the local independence assumption, the assumption that misclassification is independent of covariates and the assumption that covariates are free of error. The local independence assumption is always made when applying latent class models and therefore when applying the MILC method. This means that it is assumed that the observed indicators are independent of each other given a unit's score on the latent variable. When a mistake is made in an indicator originating from one source, this is independent

of mistakes made on indicators from other sources. A second assumption made is that misclassification is independent of covariates. In practice, situations can occur where this assumption is not met, but this could be handled in a quite straightforward manner by incorporating the relevant subgroups in the latent class model. However, the performance of the MILC method for such situations also remains to be investigated. A third assumption made is that the covariates are free of error. In many cases, this can be considered an unrealistic assumption, especially since the main aim when applying the MILC method is to correct for misclassification. However, if there is any information about the misclassification, for example because the missingness is related to an observed variable, it can be incorporated in the latent class model. Furthermore, the degree of misclassification probably also differs for different types of variables. For example in Chapter 3, MILC is applied to correct for misclassification in the variable *vehicle type*. Research has shown here that misclassification is substantive and problematic. The dataset also contains a variable measuring *gender*, which is included as a covariate in the model. In this situation, it is reasonable to assume that the amount of misclassification in *vehicle type* is of a whole different level compared to the minor random errors possibly made in the variable *gender*. The consequences of violating these assumptions when applying the MILC method should be thoroughly investigated.

Assumptions are also made when including edit restrictions in the latent class model. It is assumed that the edits applied are so-called 'hard edits', which means that a specific cell is fixed to have exactly the probability specified. The probability we used in both Chapter 3 and Chapter 7 is zero, hereby we are completely preventing impossible combinations of scores from occurring. Alternative edit restrictions that could in theory be incorporated in the latent class model are 'soft edits', where a specific small probability is assigned to a specific cell (De Waal et al., 2011, p.10). In addition, hard or soft edits could also be applied using specific known totals using macro-integration. However, direct comparisons between the MILC method and other edit correction methods are not made in this thesis.

8.2.3 Parameter uncertainty

An issue discussed within many chapters of this thesis, is parameter uncertainty. Within the MILC method, parameter uncertainty has generally been incorporated by drawing non-parametric bootstrap samples from the observed dataset. Next, a latent class model is estimated on each of the bootstrap samples instead of on the observed data directly. The differences then reflect uncertainty about the parameters. There were no issues with this bootstrap when a small set-up was used in simulation studies, as in Chapter 2 and Chapter 5.

However, when the latent class model was more complex, such as in Chapter 3 with more covariates and more categories or in Chapter 4 with multiple time-points, the use of this bootstrap was more problematic. In these situations, the latent class model consists of a larger number of parameters. If a bootstrap sample is then drawn from the original dataset, and a latent class model is estimated on that bootstrap sample, it becomes more likely that a specific parameter that is needed to create imputations for the original data is not estimated within the latent class model. It is then not possible to create an imputation for the profiles related to that parameter for that specific bootstrap. Alternatives could be a parametric bootstrap or Bayesian MCMC, but the feasibility and flexibility of implementing these within the MILC framework should be checked.

8.2.4 Total Survey Error framework

When unit linked data-sets are used to produce statistics, random misclassification (the types of errors for which we investigated the performance of MILC) is not the only source of error. Many other types of errors and sources of error can be identified. When only a survey is used to produce statistics, the Total Survey Error framework provides a systematic overview of possible errors (Groves et al., 2011). B. F. M. Bakker (2010) and Zhang (2012) extended this framework for combined survey-register datasets. The latter proposed a two-phase life cycle, where the primary cycle refers to errors located within a single source and the secondary cycle refers to errors uncovered or induced while integrating multiple sources.

This research was mainly aimed at correcting for misclassification or, in other words, measurement error. Strictly defined, measurement error is the error that occurs when the recorded or observed value is different from the true value of the variable (Alwin, 2007, p.6), and this can be distinguished from other types of errors, such as sampling error (error resulting from using a subset of the population instead of the population itself), coverage error (error due to some population elements having an inclusion probability of zero) or nonresponse error (error resulting from not obtaining data from all selected respondents, Alwin, 2007; De Leeuw et al., 2012). However, when correcting for misclassification using the MILC method, we consider by definition the situation of a combined dataset, and the two-phase life cycle of errors as defined by Zhang (2012) is applicable. A clear distinction between different types of measurement error is not made in this thesis. Therefore, it should be investigated in more detail if other types of errors can also be detected and corrected for when using the MILC method. For example, Pankowska et al. (n.d.) already showed that

when a latent Markov model is used to estimate misclassification in a dataset contaminated by not only misclassification but also linkage error, the method handles linkage error and misclassification equally. A follow-up study could identify the different types of error and evaluate if, how and when the MILC method is able to correct for each type of error. In addition, it should be investigated what correction methods are available to handle the types of error that cannot be estimated and corrected for by the MILC method, and the feasibility of integrating these methods with the MILC method should be investigated. A first example is nonresponse error. It was already shown in Chapter 3 and Chapter 4 that in some situations the MILC method is able to handle nonresponse error as well, so research could be done on how the MILC method can be extended further to for example simultaneously create multiple imputations for missing values in covariates. A second example is coverage error. It is not reasonable to think that the MILC method is able to correct for this type of error directly and capture/recapture methods are typically used to correct for this. Therefore, it should be investigated how the MILC method and capture/recapture methods can be combined. Alternatively, a latent class model could be used to trace 'duplicated scores' (resulting in over-coverage, Di Cecco et al., 2018; Steorts et al., 2016) so investigation could also be done on the integration of this coverage-model into the MILC method.

In the context of the Total Survey Error framework, it is also relevant to make a comparison between MILC and weighting. When both survey and register data are available, weighting is typically used to weight the survey data in such a way that it appropriately represents the target population, accounting for issues such as nonresponse and stratified or clustered sampling schemes (Bethlehem et al., 2011). If a latent class model is used to estimate misclassification, posterior membership probabilities can in theory also be used for weighting directly. However, this does not take into account issues such as logically impossible combinations of scores and consistency over different types of output, which can be accounted for by making use of repeated weighting (Daalmans, 2015). However, one mayor issue remains unaccounted for. When weighting a combined dataset, the weighting procedure needs to be applied and tailored for each analysis performed on a combined dataset. When different variables are used originating from different samples within the combined dataset, different sets of weights need to be used. Applied researchers need to have extensive methodological knowledge for every analysis performed on a specific combined dataset. When applying the MILC procedure, a new, multiply imputed dataset, is generated that can be corrected for all issues discussed. In this thesis, we evaluated general misclassification, misclassification resulting in logically impossible combinations of scores and unit non-response, but in theory it should also be possible to incorporate item-nonresponse and non-

random sampling schemes. Thus, the only methodological knowledge an applied researcher needs is on how to work with this multiply imputed dataset. As a result, the preparation procedure of the data can truly be separated and is completely independent of the analysis phase. In addition, the independence between these two procedures can increase objectivity when preparing the dataset and can result in increased replicatability of research in general.

To summarize, this thesis has shown that multiple imputation of latent classes is a flexible solution to simultaneously estimate and correct for misclassification and missing data in combined datasets.

Appendix A

Table A1: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of Y_1 and covariate Z with different values for the classification probabilities, different values for $P(Z = 2)$ and different values for sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	P(Z=2)	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	-.0030	.9490	1.0282	-.0020	.9490	1.0315	-.0011	.9400	1.0072	-.0009	.9460	1.0053	-.0006	.9370	0.9730
		2	.0031	.9480	1.0301	.0020	.9480	1.0312	.0011	.9390	1.0089	.0009	.9410	0.9946	.0005	.9350	0.9611
		3	.0030	.5690	0.9032	.0019	.7290	0.9144	.0010	.6370	0.8257	.0005	.3820	0.6093	.0001	.0950	0.3097
		4	-.0031	.6880	0.9381	-.0020	.8210	1.0064	-.0010	.8840	0.9526	-.0005	.9070	0.9748	-.0000	.9330	0.9967
	.05	1	-.0146	.8340	0.9803	-.0096	.9150	1.0300	-.0055	.9250	0.9860	-.0030	.9280	0.9716	-.0008	.9470	0.9994
		2	.0144	.8530	0.9799	.0097	.9120	1.0207	.0052	.9370	0.9843	.0027	.9330	0.9723	.0010	.9630	0.9980
		3	.0150	.0000	0.9944	.0099	.0120	1.0129	.0050	.2480	0.9638	.0025	.6600	0.8902	.0005	.3880	0.6134
		4	-.0147	.2740	1.0104	-.0100	.5850	0.9858	-.0047	.8560	1.0292	-.0022	.9020	0.9713	-.0007	.9300	0.9680
	.10	1	-.0297	.5130	0.9659	-.0202	.7450	1.0181	-.0093	.9010	1.0017	-.0051	.9320	0.9894	-.0004	.9540	1.0310
		2	.0297	.5210	0.9686	.0204	.7410	1.0152	.0093	.8970	0.9680	.0047	.9300	0.9626	.0002	.9520	1.0109
		3	.0300	.0000	0.9885	.0202	.0000	1.0051	.0099	.0110	0.9797	.0050	.2690	0.9833	.0010	.6070	0.7905
		4	-.0300	.0480	1.0377	-.0204	.3570	0.9469	-.0099	.8030	1.0010	-.0046	.9150	1.0553	-.0008	.9460	0.9981
	.20	1	-.0600	.0240	0.9841	-.0400	.2760	0.9638	-.0198	.7460	0.9571	-.0092	.9020	0.9908	-.0018	.9420	0.9911
		2	.0593	.0270	0.9931	.0405	.2260	0.9983	.0203	.7310	0.9611	.0090	.8990	0.9822	.0020	.9430	0.9949
		3	.0605	.0000	0.9737	.0399	.0000	0.9931	.0199	.0000	0.9830	.0100	.0100	0.9891	.0020	.7300	0.8763
		4	-.0597	.0000	1.0523	-.0404	.0710	0.9887	-.0204	.5840	0.9690	-.0098	.8610	0.9887	-.0022	.9350	0.9769
10,000	.01	1	-.0031	.9060	1.0085	-.0022	.9470	1.0526	-.0009	.9460	0.9997	-.0006	.9170	0.9558	-.0000	.9460	0.9592
		2	.0031	.9090	1.0003	.0021	.9440	1.0559	.0009	.9450	1.0010	.0006	.9310	0.9685	.0000	.9470	0.9599
		3	.0030	.0000	1.0166	.0020	.0000	0.9763	.0010	.0130	0.9765	.0005	.2980	0.9257	.0001	.6010	0.7449
		4	-.0030	.0670	1.0061	-.0020	.4020	1.0053	-.0010	.7800	0.9837	-.0005	.8990	1.0001	-.0001	.9570	1.0448
	.05	1	-.0149	.1550	0.9830	-.0102	.4580	0.9722	-.0049	.8440	1.0069	-.0024	.9360	1.0488	-.0005	.9460	0.9902
		2	.0148	.1460	0.9879	.0103	.4550	0.9905	.0048	.8380	0.9870	.0024	.9290	1.0282	.0005	.9530	1.0145
		3	.0150	.0000	0.9912	.0100	.0000	0.9905	.0050	.0000	0.9780	.0025	.0000	1.0130	.0005	.2400	1.0204
		4	-.0150	.0000	1.0053	-.0101	.0000	0.9859	-.0049	.3550	1.0221	-.0025	.7780	0.9873	-.0005	.9410	0.9776
	.10	1	-.0298	.0000	1.0235	-.0200	.0210	0.9755	-.0101	.4720	1.0048	-.0052	.8140	0.9984	-.0010	.9460	1.0379
		2	.0298	.0000	1.0135	.0202	.0160	0.9866	.0100	.4910	1.0333	.0050	.8180	0.9866	.0010	.9510	1.0240
		3	.0300	.0000	1.0127	.0199	.0000	1.0195	.0100	.0000	0.9927	.0050	.0000	0.9951	.0010	.0140	1.0053
		4	-.0300	.0000	1.0421	-.0201	.0000	1.0141	-.0100	.0640	1.0627	-.0048	.6220	1.0058	-.0010	.9400	1.0022
	.20	1	-.0600	.0000	1.0166	-.0401	.0000	1.0019	-.0202	.0230	1.0247	-.0098	.4960	1.0247	-.0023	.9160	0.9981
		2	.0599	.0000	1.0085	.0400	.0000	1.0031	.0201	.0140	0.9854	.0099	.4280	1.0534	.0021	.9180	0.9838
		3	.0600	.0000	1.0068	.0401	.0000	0.9837	.0199	.0000	1.0067	.0100	.0000	1.0267	.0020	.0000	0.9471
		4	-.0599	.0000	0.9514	-.0399	.0000	1.0110	-.0199	.0000	0.9832	-.0100	.2730	1.0208	-.0018	.9260	0.9964

Table A2: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the unconditional model and covariate Z with different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	P(Z=2)	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	.0366	.8470	1.0021	.0394	.7770	0.9261	.0173	.8320	0.9552	.0037	.9420	1.0147	.0006	.9480	1.0053
		2	-.0368	.8520	1.0019	-.0393	.7780	0.9289	-.0173	.8370	0.9612	-.0037	.9430	1.0136	-.0005	.9510	1.0140
		3	.0027	.9370	1.2328	.0010	.8400	1.2466	.0002	.3170	0.8594	.0000	.0810	0.4324	.0000	.0040	0.0974
		4	-.0025	.8160	1.1160	-.0010	.9080	1.0406	-.0002	.9130	0.9804	-.0000	.9340	1.0045	-.0000	.9240	0.9772
	.05	1	.0249	.8540	1.0331	.0333	.8040	0.9736	.0153	.8770	0.9936	.0054	.9350	0.9894	.0016	.9620	1.0258
		2	-.0248	.8540	1.0347	-.0332	.8070	0.9743	-.0152	.8630	0.9943	-.0055	.9360	0.9994	-.0012	.9600	1.0235
		3	.0131	.6080	1.1421	.0049	.8320	1.3231	.0009	.7990	1.1260	.0002	.3140	0.7553	.0000	.0160	0.1506
		4	-.0132	.6620	1.0697	-.0050	.8740	1.0218	-.0010	.9320	0.9923	-.0001	.9430	0.9812	-.0005	.9590	1.0732
	.10	1	.0102	.8460	0.9645	.0253	.8430	0.9270	.0124	.8870	0.9782	.0041	.9340	0.9633	.0006	.9460	1.0048
		2	-.0102	.8450	0.9640	-.0250	.8380	0.9310	-.0120	.8820	0.9706	-.0039	.9330	0.9830	-.0009	.9520	1.0054
		3	.0260	.4920	1.0207	.0098	.5010	1.1634	.0017	.9490	1.3165	.0004	.4960	0.9755	.0000	.0270	0.2549
		4	-.0260	.5490	1.0473	-.0100	.7800	0.9656	-.0021	.9390	1.0089	-.0005	.9350	0.9589	.0003	.9580	1.0001
	.20	1	-.0215	.8540	0.9901	.0112	.8720	0.8983	.0076	.9240	0.9714	.0022	.9440	0.9852	-.0007	.9450	1.0061
		2	.0220	.8540	0.9917	-.0107	.8700	0.9106	-.0075	.9300	0.9971	-.0022	.9430	0.9964	.0007	.9420	1.0025
		3	.0501	.4130	1.0317	.0189	.2250	1.0623	.0031	.9330	1.3470	.0006	.7070	1.1819	.0000	.0620	0.3982
		4	-.0506	.4990	1.0398	-.0194	.7040	0.9939	-.0032	.9290	0.9898	-.0005	.9590	1.0322	.0000	.9530	0.9921
10,000	.01	1	.0439	.7410	0.9406	.0420	.1180	0.9371	.0163	.2460	0.9242	.0049	.8280	0.9624	.0005	.9490	1.0112
		2	-.0440	.7390	0.9405	-.0420	.1160	0.9362	-.0163	.2560	0.9124	-.0049	.8260	0.9596	-.0005	.9480	1.0085
		3	.0025	.0170	1.2140	.0010	.2510	1.3434	.0002	.9530	1.2920	.0000	.4960	1.0528	.0000	.0200	0.2565
		4	-.0024	.3630	1.0981	-.0009	.8280	1.0447	-.0002	.9300	1.0030	-.0000	.9440	0.9849	-.0000	.9530	1.0248
	.05	1	.0314	.7980	0.9143	.0347	.1910	0.9367	.0143	.3370	0.9716	.0043	.8760	1.0095	.0001	.9450	0.9906
		2	-.0314	.7990	0.9117	-.0346	.1910	0.9476	-.0144	.3100	0.9818	-.0042	.8740	0.9970	-.0001	.9460	0.9851
		3	.0122	.0000	0.9927	.0047	.0000	1.2805	.0008	.3410	1.3722	.0001	.9330	1.3488	.0000	.1130	0.6631
		4	-.0123	.0170	1.0485	-.0047	.4330	1.0148	-.0007	.9470	1.0245	-.0002	.9540	1.0424	.0001	.9520	0.9968
	.10	1	.0172	.8690	0.9719	.0269	.3530	0.9559	.0123	.4490	0.9795	.0042	.8530	0.9596	.0001	.9440	0.9698
		2	-.0172	.8660	0.9749	-.0267	.3540	0.9753	-.0123	.4190	0.9746	-.0042	.8640	0.9899	-.0002	.9560	1.0132
		3	.0243	.0000	1.0485	.0092	.0000	1.2298	.0016	.0250	1.3530	.0003	.9630	1.4935	.0000	.1750	0.8093
		4	-.0243	.0030	1.0305	-.0094	.1400	1.0139	-.0016	.9100	1.0152	-.0003	.9580	1.0118	.0001	.9630	1.0389
	.20	1	-.0135	.8600	0.9420	.0111	.7430	0.9338	.0078	.7000	0.9701	.0025	.9220	0.9899	.0004	.9470	0.9896
		2	.0134	.8600	0.9490	-.0113	.7220	0.9505	-.0078	.6810	0.9894	-.0025	.9090	0.9983	-.0003	.9500	0.9825
		3	.0474	.0000	0.9731	.0177	.0000	1.0894	.0030	.0000	1.3102	.0005	.7840	1.4592	.0000	.2640	1.0452
		4	-.0473	.0010	0.9562	-.0175	.0180	0.9876	-.0030	.8760	0.9922	-.0005	.9470	0.9829	-.0001	.9550	1.0227

Table A3: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the conditional model and covariate Z with different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	P(Z=2)	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	-.2642	.4510	0.9192	.0544	.6940	0.9307	.0188	.8180	0.9471	.0037	.9420	0.9971	.0102	.9050	1.0478
		2	.2640	.4530	0.9210	-.0544	.6980	0.9304	-.0188	.8150	0.9519	-.0037	.9430	0.9962	-.0102	.9170	1.0569
		3	-.0025	.8190	1.1097	.0005	.5420	0.9409	.0001	.2030	0.7171	.0000	.0610	0.4060	-.0100	.0000	0.0746
		4	.0027	.9640	1.3693	-.0005	.9190	1.0043	-.0001	.9140	0.9741	-.0000	.9350	1.0036	.0100	.0100	1.0072
	.05	1	.1163	.6880	1.0093	.0543	.6570	0.9572	.0173	.8460	0.9963	.0055	.9310	0.9908	.0016	.9640	1.0268
		2	-.1162	.6890	1.0128	-.0543	.6400	0.9564	-.0171	.8420	0.9956	-.0056	.9330	0.9989	-.0011	.9620	1.0251
		3	.0055	.9170	1.0301	.0011	.7930	0.9647	.0002	.3500	0.8031	.0000	.0960	0.5438	.0000	.0150	0.1936
		4	-.0056	.8770	0.9839	-.0012	.9370	0.9854	-.0003	.9410	0.9926	.0000	.9430	0.9793	-.0005	.9590	1.0746
	.10	1	.1040	.6550	1.0122	.0474	.6600	0.9330	.0144	.8740	0.9752	.0042	.9320	0.9683	.0006	.9480	1.0060
		2	-.1040	.6480	1.0132	-.0472	.6510	0.9297	-.0140	.8610	0.9594	-.0041	.9360	0.9862	-.0009	.9560	1.0074
		3	.0084	.9190	0.9661	.0018	.8420	0.9172	.0002	.3530	0.7467	.0000	.0650	0.4029	.0000	.0160	0.2591
		4	-.0084	.8970	1.0497	-.0021	.9250	0.9437	-.0006	.9590	1.0080	-.0002	.9470	0.9595	.0002	.9570	1.0004
	.20	1	.0723	.6580	0.9574	.0354	.7010	0.9143	.0100	.9090	0.9647	.0024	.9450	0.9822	-.0009	.9440	1.0038
		2	-.0718	.6520	0.9722	-.0349	.6840	0.9238	-.0099	.9050	0.9785	-.0024	.9420	0.9975	.0008	.9420	1.0013
		3	.0118	.9290	0.9710	.0026	.9150	0.9981	.0003	.3880	0.7857	.0000	.0390	0.2919	.0000	.0250	0.3337
		4	-.0123	.9050	1.0121	-.0031	.9480	1.0107	-.0004	.9460	0.9904	.0000	.9580	1.0282	-.0000	.9550	0.9929
10,000	.01	1	.1083	.3180	0.9304	.0580	.0140	0.9355	.0177	.1860	0.9253	.0050	.8180	0.9618	.0005	.9480	1.0122
		2	-.1084	.3150	0.9299	-.0581	.0140	0.9341	-.0178	.1750	0.9133	-.0050	.8140	0.9592	-.0005	.9480	1.0094
		3	.0006	.9300	0.9719	.0002	.8490	0.9536	.0000	.3510	0.7248	.0000	.1330	0.5662	.0000	.0180	0.2357
		4	-.0006	.9140	1.0145	-.0001	.9510	1.0016	-.0000	.9330	0.9943	-.0000	.9440	0.9827	-.0000	.9530	1.0248
	.05	1	.1036	.1730	0.9255	.0524	.0130	0.9195	.0159	.2330	0.9716	.0044	.8530	1.0045	.0001	.9450	0.9891
		2	-.1036	.1710	0.9222	-.0524	.0110	0.9285	-.0160	.2270	0.9780	-.0044	.8620	0.9924	-.0001	.9470	0.9840
		3	.0018	.8800	0.9726	.0004	.9500	0.9899	.0001	.5210	0.8260	.0000	.1170	0.4172	.0000	.0230	0.3223
		4	-.0018	.9020	1.0385	-.0005	.9430	0.9846	.0001	.9600	1.0233	-.0001	.9590	1.0413	.0001	.9520	0.9969
	.10	1	.0896	.0740	0.9529	.0455	.0180	0.9337	.0140	.3380	0.9850	.0043	.8500	0.9594	.0001	.9440	0.9693
		2	-.0897	.0730	0.9515	-.0453	.0170	0.9555	-.0140	.3200	0.9809	-.0043	.8570	0.9918	-.0002	.9560	1.0124
		3	.0026	.8500	0.9642	.0007	.9210	0.9950	.0001	.6890	0.8811	.0000	.1460	0.4519	.0000	.0080	0.1825
		4	-.0025	.9170	1.0370	-.0008	.9310	0.9986	-.0001	.9560	1.0123	-.0000	.9560	1.0116	.0001	.9640	1.0389
	.20	1	.0601	.0970	0.9126	.0312	.0760	0.9378	.0098	.5890	0.9694	.0027	.9240	0.9922	.0004	.9470	0.9894
		2	-.0602	.0910	0.9402	-.0313	.0570	0.9654	-.0097	.5310	0.9889	-.0027	.9080	1.0019	-.0003	.9510	0.9822
		3	.0036	.8420	0.9642	.0009	.9090	0.9216	.0002	.8180	1.0068	.0000	.2220	0.5693	.0000	.0020	0.0715
		4	-.0035	.8960	1.0040	-.0007	.9330	0.9912	-.0002	.9500	0.9984	-.0000	.9480	0.9811	-.0001	.9550	1.0229

Table A4: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the restricted conditional model and covariate Z with different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	P(Z=2)	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	.0920	.7870	1.0070	.0561	.6670	0.9301	.0190	.8210	0.9555	.0038	.9410	1.0149	-.0003	.9360	0.9726
		2	-.0922	.7880	1.0075	-.0561	.6630	0.9322	-.0190	.8220	0.9608	-.0038	.9420	1.0149	.0002	.9400	0.9595
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0002	.9330	0.9656	-.0001	.9290	1.0017	.0000	.9140	0.9738	.0000	.9360	1.0045	.0001	.9380	1.0022
	.05	1	.0780	.7740	0.9893	.0507	.6730	0.9450	.0171	.8520	0.9882	.0056	.9330	0.9864	.0016	.9630	1.0270
		2	-.0779	.7690	0.9927	-.0507	.6630	0.9456	-.0170	.8460	0.9896	-.0056	.9320	0.9962	-.0011	.9610	1.0249
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	-.0001	.9200	0.9375	-.0001	.9360	0.9757	-.0001	.9410	0.9919	.0001	.9440	0.9800	-.0005	.9590	1.0741
	.10	1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663	.0006	.9480	1.0044
		2	-.0705	.7450	0.9894	-.0423	.6770	0.9283	-.0137	.8760	0.9607	-.0041	.9320	0.9822	-.0009	.9540	1.0052
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	-.0000	.9510	0.9994	-.0003	.9380	0.9340	-.0004	.9600	1.0104	-.0002	.9470	0.9589	.0003	.9580	0.9994
	.20	1	.0501	.7530	0.9482	.0309	.7440	0.9212	.0095	.9060	0.9686	.0024	.9410	0.9801	-.0007	.9470	1.0041
		2	-.0496	.7500	0.9611	-.0304	.7330	0.9279	-.0094	.9040	0.9845	-.0024	.9390	0.9907	.0007	.9430	1.0000
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	-.0005	.9490	1.0238	-.0005	.9530	1.0044	-.0001	.9460	0.9915	.0001	.9560	1.0284	.0000	.9530	0.9923
10,000	.01	1	.1021	.3450	0.9331	.0581	.0110	0.9398	.0178	.1780	0.9292	.0050	.8200	0.9664	.0005	.9500	1.0132
		2	-.1021	.3440	0.9330	-.0582	.0120	0.9391	-.0178	.1820	0.9168	-.0050	.8220	0.9633	-.0005	.9480	1.0101
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9470	0.9863	.0001	.9510	1.0032	.0000	.9310	0.9913	-.0000	.9440	0.9825	-.0000	.9530	1.0248
	.05	1	.0916	.2140	0.9175	.0513	.0140	0.9214	.0158	.2380	0.9697	.0044	.8630	1.0099	.0001	.9460	0.9894
		2	-.0916	.2120	0.9145	-.0512	.0130	0.9309	-.0159	.2170	0.9757	-.0044	.8690	0.9979	-.0001	.9470	0.9840
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	-.0000	.9530	0.9966	-.0001	.9490	0.9824	.0001	.9600	1.0234	-.0001	.9590	1.0416	.0001	.9520	0.9971
	.10	1	.0794	.1050	0.9382	.0441	.0240	0.9314	.0139	.3470	0.9666	.0043	.8510	0.9603	.0001	.9460	0.9697
		2	-.0794	.1050	0.9361	-.0439	.0180	0.9537	-.0139	.3200	0.9653	-.0043	.8520	0.9925	-.0002	.9560	1.0129
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0000	.9580	1.0304	-.0002	.9390	0.9953	-.0000	.9560	1.0115	-.0000	.9560	1.0118	.0001	.9640	1.0389
	.20	1	.0531	.1680	0.9064	.0300	.0890	0.9311	.0096	.6020	0.9619	.0027	.9250	0.9907	.0004	.9450	0.9898
		2	-.0532	.1530	0.9333	-.0301	.0620	0.9565	-.0096	.5350	0.9772	-.0026	.9130	0.9982	-.0003	.9510	0.9831
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9460	0.9888	.0002	.9470	0.9933	-.0001	.9490	0.9993	-.0000	.9480	0.9803	-.0001	.9550	1.0229

Table A5: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of Y_1 regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	P(Z=2)	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	-.1139	.8550	1.0087	-.0763	.9080	1.0065	-.0432	.9430	0.9924	-.0253	.9360	0.9852	-.0047	.9590	1.0096
		.05	-.1254	.8360	1.0070	-.0756	.9220	1.0013	-.0503	.9230	0.9830	-.0170	.9290	0.9661	-.0025	.9500	0.9876
		.10	-.1156	.8620	1.0103	-.0823	.8970	0.9963	-.0335	.9550	1.0339	-.0228	.9420	0.9940	-.0066	.9460	0.9528
		.20	-.1144	.8610	1.0121	-.0762	.9030	0.9795	-.0378	.9430	1.0232	-.0157	.9500	1.0179	-.0063	.9480	0.9803
	.55	.01	.1221	.8440	1.0323	.0808	.8930	0.9760	.0378	.9440	1.0077	.0204	.9460	0.9819	.0040	.9450	0.9799
		.05	.1156	.8590	1.0206	.0795	.9000	0.9768	.0329	.9400	0.9950	.0217	.9320	0.9681	.0027	.9560	1.0533
		.10	.1250	.8440	1.0136	.0753	.9050	1.0200	.0376	.9310	0.9820	.0214	.9400	1.0049	.0080	.9450	0.9828
		.20	.1164	.8530	1.0150	.0759	.9000	0.9940	.0403	.9350	0.9989	.0178	.9410	0.9894	.0052	.9580	1.0157
	.65	.01	.3768	.1790	0.9792	.2498	.5060	1.0233	.1322	.8310	0.9955	.0596	.9270	0.9876	.0135	.9480	1.0014
		.05	.3677	.1700	1.0074	.2448	.5270	0.9782	.1249	.8260	0.9749	.0604	.9210	1.0037	.0021	.9440	0.9610
		.10	.3700	.1660	1.0439	.2456	.5260	1.0248	.1248	.8500	1.0019	.0659	.9230	1.0028	.0018	.9450	0.9874
		.20	.3799	.1570	1.0223	.2557	.4880	0.9879	.1258	.8490	1.0087	.0648	.9290	1.0449	.0173	.9440	0.9946
10,000	.45	.01	-.1211	.1510	0.9838	-.0794	.5020	0.9382	-.0417	.8140	1.0021	-.0222	.9240	1.0272	-.0014	.9460	0.9907
		.05	-.1201	.1510	0.9765	-.0810	.4700	1.0168	-.0399	.8320	1.0117	-.0183	.9190	0.9928	-.0018	.9440	0.9919
		.10	-.1200	.1490	0.9682	-.0813	.4550	1.0213	-.0398	.8330	1.0032	-.0199	.9100	0.9696	-.0028	.9550	1.0037
		.20	-.1179	.1730	0.9627	-.0793	.4890	0.9764	-.0411	.8190	0.9954	-.0199	.9330	1.0578	-.0029	.9550	1.0297
	.55	.01	.1195	.1470	1.0344	.0802	.4830	0.9700	.0383	.8320	1.0036	.0200	.9150	0.9717	.0058	.9470	0.9691
		.05	.1211	.1400	0.9898	.0800	.4870	0.9555	.0407	.8260	1.0384	.0193	.9200	1.0074	.0060	.9370	0.9844
		.10	.1221	.1420	0.9849	.0814	.4810	0.9800	.0373	.8600	1.0222	.0180	.9310	1.0053	.0010	.9610	1.0464
		.20	.1213	.1360	1.0191	.0786	.5080	1.0016	.0413	.8110	0.9906	.0214	.9160	0.9963	.0027	.9530	1.0121
	.65	.01	.3727	.0000	0.9855	.2499	.0000	0.9948	.1251	.1420	1.0078	.0617	.6740	0.9620	.0104	.9520	1.0347
		.05	.3735	.0000	1.0096	.2484	.0000	1.0109	.1270	.1300	0.9911	.0649	.6380	0.9782	.0145	.9310	1.0156
		.10	.3746	.0000	0.9988	.2493	.0000	1.0022	.1241	.1580	0.9729	.0624	.6740	1.0216	.0115	.9360	0.9955
		.20	.3706	.0000	1.0016	.2510	.0000	0.9942	.1266	.1250	1.0222	.0642	.6440	0.9656	.0115	.9450	1.0145

Table A6: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the unconditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	P(Z=2)	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	.0157	.9350	0.9918	.0090	.9530	1.0346	-.0021	.9520	1.0093	.0015	.9460	0.9795	-.0053	.9500	1.0045
		.05	.0081	.9410	1.0813	-.0010	.9320	0.9798	.0024	.9490	1.0050	-.0047	.9530	1.0270	.0029	.9510	0.9889
	.45	.10	.0167	.9310	0.9873	-.0028	.9410	0.9611	.0050	.9450	0.9953	.0011	.9580	1.0166	.0016	.9460	0.9968
		.20	.0112	.9210	0.9989	.0079	.9400	0.9890	.0006	.9520	0.9672	-.0046	.9580	1.0020	-.0005	.9520	1.0285
		.01	-.0308	.9310	0.9980	-.0031	.9350	0.9906	-.0006	.9440	0.9985	-.0093	.9470	0.9968	.0084	.9520	1.0232
		.05	-.0090	.9230	0.9695	-.0047	.9350	0.9788	-.0006	.9540	0.9978	-.0043	.9520	1.0365	-.0002	.9610	1.0240
1,000	.55	.10	-.0214	.9310	0.9734	-.0039	.9440	1.0123	-.0013	.9420	0.9833	.0019	.9450	0.9809	-.0045	.9600	1.0210
		.20	-.0278	.9360	0.9902	-.0112	.9460	1.0053	.0018	.9510	1.0056	.0049	.9560	1.0109	.0006	.9560	1.0235
		.01	-.0345	.9310	1.0286	-.0049	.9470	1.0130	-.0013	.9360	0.9722	-.0071	.9500	1.0091	-.0009	.9490	0.9949
		.05	-.0444	.9320	1.0059	-.0098	.9400	0.9911	-.0001	.9510	1.0101	.0024	.9580	1.0171	.0028	.9420	0.9774
	.65	.10	-.0578	.9190	0.9891	-.0149	.9470	1.0139	-.0062	.9590	0.9939	-.0091	.9430	0.9583	-.0054	.9530	0.9910
		.20	-.0361	.9460	1.0222	-.0051	.9430	1.0044	-.0044	.9600	1.0294	-.0080	.9510	1.0019	-.0048	.9410	0.9982
		.01	-.0046	.9280	0.9998	.0031	.9410	0.9762	.0024	.9530	1.0250	-.0022	.9550	1.0359	.0006	.9510	0.9851
		.05	-.0020	.9210	0.9806	.0005	.9500	1.0033	.0016	.9410	1.0022	.0021	.9380	0.9844	.0014	.9500	0.9966
	.45	.10	.0044	.9200	0.9365	-.0017	.9520	1.0266	.0018	.9670	1.0462	-.0010	.9510	1.0170	.0018	.9510	1.0040
		.20	.0038	.9110	0.9531	.0012	.9510	0.9738	.0002	.9560	1.0014	-.0003	.9680	1.0423	-.0019	.9400	0.9699
		.01	-.0043	.9150	0.9728	-.0009	.9350	0.9699	-.0019	.9360	0.9678	-.0000	.9510	0.9889	-.0002	.9410	0.9808
		.05	-.0034	.9320	0.9775	-.0007	.9320	0.9671	-.0003	.9480	1.0049	-.0004	.9550	1.0133	.0019	.9500	0.9753
10,000	.55	.10	-.0080	.9310	0.9924	-.0009	.9530	1.0127	-.0033	.9430	0.9760	.0003	.9510	0.9899	-.0005	.9450	0.9955
		.20	-.0069	.9210	0.9813	-.0028	.9460	1.0044	-.0010	.9340	0.9476	.0008	.9550	1.0331	.0001	.9490	1.0081
		.01	-.0083	.9250	0.9643	-.0014	.9300	0.9458	-.0037	.9530	1.0423	-.0033	.9500	0.9952	-.0001	.9630	1.0809
		.05	-.0062	.9260	1.0011	-.0024	.9390	0.9908	-.0029	.9520	0.9845	.0002	.9410	0.9877	.0013	.9620	1.0461
	.65	.10	-.0153	.9300	0.9918	-.0053	.9440	0.9832	-.0028	.9570	1.0140	-.0013	.9530	1.0042	.0008	.9570	1.0047
		.20	-.0134	.9030	0.9355	-.0032	.9370	0.9849	.0001	.9470	0.9967	-.0005	.9490	1.0010	-.0001	.9610	1.0455

Table A7: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	P(Z=2)	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	-.4303	.7100	1.3597	.0097	.9550	1.0372	-.0036	.9580	1.0172	.0022	.9450	0.9789	-.0053	.9470	1.0040
		.05	.0075	.9490	1.0825	-.0027	.9360	0.9812	.0025	.9520	0.9975	-.0043	.9550	1.0289	.0032	.9500	0.9899
	.45	.10	.0135	.9310	0.9973	-.0021	.9410	0.9624	.0065	.9440	0.9964	.0014	.9570	1.0134	.0016	.9480	0.9982
		.20	.0027	.9290	0.9731	.0027	.9410	1.0048	.0003	.9540	0.9760	-.0048	.9470	0.9990	-.0006	.9540	1.0283
		.01	.4050	.7500	1.2376	-.0039	.9360	0.9990	-.0015	.9470	1.0016	-.0096	.9490	0.9959	.0084	.9520	1.0231
		.05	-.0028	.9220	1.0031	-.0041	.9450	0.9927	-.0008	.9490	0.9984	-.0037	.9540	1.0433	-.0001	.9590	1.0239
1,000	.55	.10	-.0113	.9490	0.9894	-.0018	.9470	1.0198	-.0020	.9470	0.9814	.0025	.9460	0.9805	-.0044	.9600	1.0211
		.20	-.0087	.9450	1.0109	-.0080	.9490	1.0205	.0015	.9490	0.9990	.0056	.9590	1.0146	.0006	.9560	1.0236
		.01	1.3336	.0800	1.3089	-.0060	.9480	1.0146	-.0027	.9440	0.9774	-.0076	.9520	1.0107	-.0010	.9540	0.9959
		.05	-.0381	.9430	1.6091	-.0101	.9370	0.9926	-.0012	.9540	1.0142	.0018	.9560	1.0148	.0028	.9420	0.9773
	.65	.10	-.0316	.9310	1.0027	-.0114	.9460	1.0251	-.0070	.9590	0.9969	-.0089	.9340	0.9574	-.0052	.9520	0.9904
		.20	-.0004	.9480	1.0298	.0040	.9450	0.9902	-.0037	.9570	1.0298	-.0082	.9560	1.0068	-.0048	.9430	0.9975
		.01	.0027	.9240	0.9560	.0031	.9390	0.9819	-.0003	.9450	0.9913	-.0009	.9510	1.0202	.0030	.9410	0.9946
		.05	.0011	.9400	1.0213	.0005	.9490	1.0116	.0021	.9560	1.0079	.0020	.9420	0.9893	.0024	.9450	0.9938
	.45	.10	.0004	.9220	0.9820	-.0026	.9540	1.0325	.0004	.9530	1.0206	.0008	.9430	0.9837	.0015	.9600	0.9878
		.20	.0006	.9380	0.9926	-.0005	.9510	0.9893	-.0019	.9550	1.0067	.0009	.9590	1.0582	.0005	.9640	1.0294
		.01	-.0070	.9220	0.9893	-.0012	.9350	0.9709	-.0018	.9440	0.9918	-.0009	.9470	0.9815	.0018	.9470	0.9713
		.05	.0012	.9200	0.9515	-.0003	.9320	0.9704	.0005	.9520	0.9998	-.0012	.9570	1.0107	.0018	.9420	0.9900
10,000	.55	.10	-.0001	.9260	0.9837	-.0001	.9570	1.0173	-.0044	.9530	1.0407	-.0028	.9480	1.0010	-.0025	.9580	1.0482
		.20	.0029	.9290	0.9886	-.0011	.9430	1.0107	-.0001	.9410	0.9692	.0013	.9520	0.9884	-.0009	.9550	1.0200
		.01	-.0105	.9220	1.1451	-.0026	.9300	0.9572	-.0044	.9580	1.0195	-.0029	.9530	0.9930	-.0022	.9540	1.0358
		.05	-.0074	.9280	0.9838	-.0025	.9390	0.9986	-.0011	.9500	0.9978	.0004	.9370	0.9773	.0014	.9540	1.0102
	.65	.10	-.0014	.9400	0.9904	-.0034	.9410	0.9913	-.0043	.9500	1.0043	-.0017	.9490	1.0400	-.0009	.9530	0.9892
		.20	.0017	.9330	1.0127	.0023	.9470	0.9892	-.0024	.9480	0.9917	.0002	.9450	0.9932	-.0011	.9600	1.0095

Table A8: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the restricted conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for $P(Z = 2)$ and sample size (N), number of bootstrap samples $m = 5$.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	P(Z=2)	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	.0318	.9170	0.9730	.0099	.9470	1.0215	.0018	.9530	0.9893	-.0054	.9400	0.9693	-.0013	.9500	1.0134
		.05	.0099	.9390	1.0035	-.0011	.9380	0.9586	-.0096	.9410	0.9932	.0077	.9490	0.9912	.0024	.9500	0.9816
	.45	.10	.0057	.9370	1.0108	-.0027	.9420	0.9728	.0047	.9580	1.0362	-.0029	.9520	1.0184	-.0033	.9430	0.9480
		.20	.0060	.9350	1.0079	.0050	.9530	1.0086	.0063	.9550	1.0681	.0051	.9530	1.0049	-.0008	.9450	0.9851
		.01	-.0088	.9370	1.0423	-.0040	.9380	0.9916	-.0055	.9540	1.0004	-.0011	.9470	0.9818	-.0002	.9480	0.9880
		.05	-.0048	.9400	1.0348	-.0037	.9460	0.9922	-.0094	.9480	1.0012	.0016	.9430	0.9808	-.0014	.9550	1.0565
1,000	.55	.10	.0058	.9340	1.0123	-.0021	.9350	0.9904	-.0011	.9460	0.9986	.0031	.9510	1.0202	.0023	.9530	0.9840
		.20	-.0010	.9320	0.9809	-.0065	.9470	1.0168	-.0019	.9510	0.9913	-.0015	.9490	0.9833	.0009	.9550	1.0090
		.01	-.0351	.9450	1.0301	-.0061	.9480	1.0107	.0008	.9450	0.9898	-.0086	.9530	1.0075	.0004	.9540	1.0022
		.05	-.0192	.9410	0.9847	-.0044	.9460	1.0003	-.0059	.9430	0.9849	-.0033	.9540	1.0015	-.0108	.9460	0.9629
	.65	.10	-.0080	.9340	1.0162	-.0139	.9480	1.0160	-.0038	.9580	1.0093	-.0006	.9580	0.9895	-.0109	.9430	0.9965
		.20	.0016	.9460	0.9996	.0044	.9490	0.9849	.0044	.9550	1.0192	-.0029	.9600	1.0578	.0045	.9480	0.9885
		.01	-.0039	.9280	1.0269	.0033	.9340	0.9739	-.0005	.9470	0.9897	-.0005	.9540	1.0205	.0030	.9410	0.9939
		.05	-.0019	.9260	0.9895	.0006	.9510	1.0142	.0020	.9500	1.0154	.0019	.9400	0.9886	.0024	.9430	0.9934
	.45	.10	-.0006	.9190	0.9415	-.0024	.9510	1.0333	.0003	.9520	1.0188	.0006	.9400	0.9867	.0015	.9580	0.9889
		.20	-.0004	.9370	0.9883	-.0005	.9550	1.0033	-.0019	.9510	1.0037	.0011	.9610	1.0530	.0006	.9620	1.0290
		.01	-.0065	.9230	0.9828	-.0015	.9380	0.9802	-.0021	.9500	0.9876	-.0007	.9490	0.9789	.0019	.9450	0.9721
		.05	.0032	.9130	0.9552	.0003	.9380	0.9804	-.0000	.9490	1.0015	-.0011	.9590	1.0104	.0018	.9420	0.9893
10,000	.55	.10	.0008	.9370	0.9855	.0006	.9520	1.0206	-.0046	.9540	1.0307	-.0030	.9490	1.0022	-.0024	.9580	1.0489
		.20	.0041	.9360	0.9792	-.0007	.9490	1.0056	.0002	.9380	0.9725	.0012	.9510	0.9893	-.0009	.9560	1.0200
		.01	-.0136	.9270	0.9887	-.0031	.9310	0.9522	-.0040	.9530	1.0141	-.0028	.9500	0.9954	-.0022	.9540	1.0352
		.05	-.0047	.9320	0.9980	-.0024	.9440	1.0046	-.0009	.9550	0.9956	.0006	.9360	0.9753	.0014	.9520	1.0110
	.65	.10	.0016	.9370	0.9886	-.0028	.9460	0.9957	-.0038	.9460	0.9928	-.0018	.9510	1.0382	-.0009	.9510	0.9892
		.20	.0058	.9370	1.0049	.0029	.9460	0.9893	-.0022	.9510	0.9946	.0001	.9460	0.9941	-.0011	.9600	1.0092

Table A9: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the restricted conditional model and covariate Z with classification probabilities 0.90, $P(Z = 2) = 0.1$, sample size=1000, 00 and different values for the number of bootstrap samples m.

		class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95				class. prob. 0.99		
m	θ	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	
	1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663		.0006	.9480	1.0044
	2	-.0705	.7450	0.9894	-.0423	.6770	0.9283	-.0137	.8760	0.9607	-.0041	.9320	0.9822		-.0009	.9540	1.0052
5	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-		.0000	-	-
	4	-.0000	.9510	0.9994	-.0003	.9380	0.9340	-.0004	.9600	1.0104	-.0002	.9470	0.9589		.0003	.9580	0.9994
	1	.07146	.7750	0.9870	.0427	.7090	0.9399	.014	.8780	0.9829	.0042	.9340	0.9690		.0006	.9460	1.0042
	2	-.0715	.7660	0.9929	-.0424	.6970	0.9406	-.0137	.8600	0.9685	-.0040	.9390	0.9861		-.0009	.9540	1.0056
10	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-		.0000	-	-
	4	.0000	.9510	0.9994	-.0003	.9380	0.9340	-.0004	.9600	1.0104	-.0002	.9470	0.9589		.0003	.9580	0.9994
	1	.0710	.8010	1.0161	.0425	.7090	0.9360	.0138	.8880	0.9753	.0041	.9300	0.9646		.0006	.9480	1.0052
	2	-.0710	.8060	1.0222	-.0423	.6930	0.9399	-.0135	.8680	0.9626	-.0040	.9340	0.9828		-.0009	.9540	1.0061
20	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-		.0000	-	-
	4	.0000	.9510	0.9994	-.0003	.9380	0.9340	-.0004	.9600	1.0104	-.0002	.9470	0.9589		.0003	.9580	0.9994
	1	0.0720	.7910	1.0152	.0426	.7120	0.9485	.0141	.8720	0.9826	.0042	.9330	0.9661		.0007	.9480	1.0045
	2	-0.0720	.7920	1.0208	-.0424	.7000	0.9512	-.0137	.8610	0.9692	-.0040	.9400	0.9835		-.0009	.9520	1.0055
40	3	0.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-		.0000	-	-
	4	-0.0000	.9510	0.9994	-.0003	.9380	0.9340	-.0004	.9600	1.0104	-.0002	.9470	0.9589		.0003	.9580	0.9994

Table A10: Entropy R^2, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the *rent benefit* variable takes information of 779 individuals into account and *marital status* variable of 3,011.

			unconditional model	conditional model	restricted conditional model
entropy R^2			0.9334	0.9377	0.9380
class. prob.	background survey	P(rent|LC rent)	0.8937	0.8938	0.9344
		P(own|LC own)	0.9997	0.9997	0.9992
	register	P(rent|LC rent)	0.9501	0.9500	0.9496
		P(own|LC own)	0.9749	0.9749	0.9525
P(rent benefit)				0.3004	0.3004
P(married)			0.5284	0.5284	0.5284

Table A11: The first block represents the (pooled) marginal proportions of the variable *own/rent*. The second block represents the (pooled) proportions of the variable *own/rent* for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable *own/rent* for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	estimate	95% CI	estimate	95% CI
register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
background survey	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
unconditional	0.6405	[0.6404; 0.6407]	0.3595	[0.3593; 0.3596]
conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
restricted conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]

	P(own; rent benefit)		P(rent; rent benefit)	
	estimate	95% CI	estimate	95% CI
register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
background survey	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
unconditional	0.0028	[0.0023; 0.0034]	0.2950	[0.2944; 0.2955]
conditional	0.0064	[-0.0263; 0.0392]	0.2914	[0.2587; 0.3241]
restricted conditional	0.0000	-	0.2978	[0.2649; 0.3307]

	P(own; no rent benefit)		P(rent; no rent benefit)	
	estimate	95% CI	estimate	95% CI
register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
background survey	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
unconditional	0.0157	[0.0151; 0.0162]	0.6829	[0.6824; 0.6835]
conditional	0.0159	[-0.0168; 0.0487]	0.6827	[0.6499; 0.7154]
restricted conditional	0.0213	[-0.0116; 0.0542]	0.6773	[0.6444; 0.7102]

Table A12: The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable *owning/renting* a house.

	intercept		marriage	
	estimate	95% CI	estimate	95% CI
register	2.4661	[2.2090; 2.7233]	-1.2331	[-1.3901; -1.0760]
background survey	2.7620	[2.4896; 3.0343]	-1.3041	[-1.4678; -1.1405]
unconditional model	2.6869	[2.4251; 2.9487]	-1.3875	[-1.6493; -1.1257]
conditional model	2.7698	[2.5034; 3.0363]	-1.3982	[-1.6646; -1.1317]
restricted conditional model	2.7712	[2.5036; 3.0389]	-1.3817	[-1.6493; -1.1140]

Appendix B

```
options
  maxthreads=all;
algorithm
  tolerance=1e-008 emtolerance=0.01 emiterations=20000  nriterations=0;
startvalues
  seed=0 sets=200 tolerance=1e-005 iterations=500;
bayes
  categorical=1 variances=1 latent=1 poisson=1;
missing
  includeall;
output
  profile;
outfile
  'posteriors1.dat' classification
keep
  LRM2, BRON2, wfactor;
variables
  caseweight b1;
  dependent LRM nominal 7, BRON nominal 7, prov_hosp nominal 12,
  prov_acc nominal 12;
  independent ernst nominal, external nominal, gender nominal,
  age nominal;
  latent X nominal 7, Xacc nominal 12;
equations
  LRM       <- 1 | X;
  BRON      <- 1 | X;
  prov_acc  <- (a~wei)Xacc;
  prov_hosp <- 1 | Xacc;
  X         <- 1 | ernst + external + gender + age;
  Xacc      <- 1 | ernst + external + gender + age;
  X <-> Xacc;
a={1 0 0 0 0 0 0 0 0 0 0 0
   0 1 0 0 0 0 0 0 0 0 0 0
   0 0 1 0 0 0 0 0 0 0 0 0
   0 0 0 1 0 0 0 0 0 0 0 0
```

```
0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 1};
```

To ensure convergence and to minimize the probability of obtaining local maxima, the number of random start sets is set to 200 with 500 iterations each. The use of Newton Rapson iterations is suppressed and the number of EM iterations is increased to $20,000$, following the suggestions by Vermunt et al. (2008).

To reduce computation time, the storing of parameters and the computation of standard errors is suppressed, since conditional and posterior response probabilities are of main interest.

To ensure that in the latent variable 'Accident of region' ($Xacc$ in the Latent GOLD syntax) the value observed in the indicator variable 'Accident of region' ($prov\_acc$ in the Latent GOLD syntax) is assigned in cases where this variable is observed, the relationship between $Xacc$ and $prov\_acc$ is restricted using the matrix denoted by 'a' in the Latent GOLD syntax.

Appendix C

Table C1: Results for X in terms of five different performance measures and obtained after using six different strategies to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0111	89.6	0.1026	1.1479	0.0253
+SI	0.0112	56.4	0.0430	0.4695	0.0259
+MI; 5	0.0112	61.2	0.0464	0.5128	0.0257
+MI; 10	0.0112	60.4	0.0461	0.5120	0.0255
+MI-B; 5	0.0221	88.8	0.1055	1.7036	0.0277
+MI-B; 10	0.0225	92.8	0.1031	1.8351	0.0269
X=2					
HMM	0.0026	94.4	0.0952	1.1662	0.0210
+SI	0.0027	65.4	0.0403	0.4821	0.0215
+MI; 5	0.0025	67.8	0.0435	0.5252	0.0213
+MI; 10	0.0025	67.2	0.0432	0.5236	0.0212
+MI-B; 5	-0.0002	99.0	0.0974	1.7733	0.0149
+MI-B; 10	-0.0006	100.0	0.0956	1.9725	0.0127
X=3					
HMM	-0.0137	85.6	0.1021	1.1644	0.0262
+SI	-0.0138	35.0	0.0265	0.2982	0.0266
+MI; 5	-0.0137	42.6	0.0315	0.3587	0.0264
+MI; 10	-0.0137	41.8	0.0312	0.3549	0.0264
+MI-B; 5	-0.0218	85.8	0.1039	1.6933	0.0275
+MI-B; 10	-0.0219	90.8	0.1022	1.8751	0.0263

Table C2: Results for X in terms of five different performance measures and obtained after using six different strategies to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.80 for both indicators

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0115	84.0	0.0743	1.0565	0.0213
+SI	0.0116	66.0	0.0431	0.5985	0.0213
+MI; 5	0.0116	69.4	0.0454	0.6431	0.0212
+MI; 10	0.0117	69.4	0.0452	0.6406	0.0212
+MI-B; 5	0.0138	89.2	0.0757	1.6893	0.0183
+MI-B; 10	0.0133	93.6	0.0753	1.8362	0.0171
X=2					
HMM	0.0022	95.0	0.0758	1.1165	0.0174
+SI	0.0022	72.8	0.0403	0.5896	0.0172
+MI; 5	0.0021	76.8	0.0427	0.6299	0.0172
+MI; 10	0.0021	76.4	0.0424	0.6269	0.0171
+MI-B; 5	-0.0006	96.4	0.0730	1.7380	0.0114
+MI-B; 10	-0.0003	99.2	0.0714	1.9208	0.0098
X=3					
HMM	-0.0138	79.8	0.0624	0.9748	0.0213
+SI	-0.0138	43.8	0.0266	0.4075	0.0189
+MI; 5	-0.0137	50.4	0.0305	0.4760	0.0189
+MI; 10	-0.0137	48.6	0.0301	0.4705	0.0189
+MI-B; 5	-0.0132	84.6	0.0635	1.4817	0.0176
+MI-B; 10	-0.0130	90.6	0.0636	1.5700	0.0169

Table C3: Results for X in terms of five different performance measures and obtained after using six different strategies to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.95 for both indicators

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0116	83.8	0.0742	1.0552	0.0214
+SI	0.0116	66.0	0.0431	0.5985	0.0214
+MI; 5	0.0116	69.4	0.0454	0.6431	0.0214
+MI; 10	0.0117	69.4	0.0452	0.6406	0.0214
+MI-B; 5	0.0183	93.6	0.0839	1.7919	0.0220
+MI-B; 10	0.0180	95.0	0.0828	1.9285	0.0212
X=2					
HMM	0.0022	95.0	0.0758	1.1146	0.0175
+SI	0.0022	72.8	0.0403	0.5896	0.0176
+MI; 5	0.0021	76.8	0.0427	0.6299	0.0174
+MI; 10	0.0021	76.4	0.0424	0.6269	0.0174
+MI-B; 5	0.0018	99.2	0.0826	1.9118	0.0116
+MI-B; 10	0.0022	100.0	0.0823	2.1573	0.0102
X=3					
HMM	-0.0138	79.8	0.0624	0.9751	0.0213
+SI	-0.0138	43.8	0.0266	0.4075	0.0217
+MI; 5	-0.0137	50.4	0.0305	0.4760	0.0214
+MI; 10	-0.0137	48.6	0.0301	0.4705	0.0214
+MI-B; 5	-0.0201	78.6	0.0691	1.4703	0.0238
+MI-B; 10	-0.0202	82.6	0.0681	1.5529	0.0233

Table C4: Results for X in terms of five different performance measures and obtained after using six different strategies to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0074	86.0	0.0597	1.0515	0.0163
+SI	0.0073	42.6	0.0193	0.3389	0.0162
+MI; 5	0.0073	45.6	0.0203	0.3568	0.0163
+MI; 10	0.0074	45.0	0.0202	0.3555	0.0163
+MI-B; 5	0.0095	94.4	0.0640	1.7185	0.0138
+MI-B; 10	0.0097	97.0	0.0621	1.8288	0.0132
X=2					
HMM	0.0016	92.2	0.0646	1.0434	0.0159
+SI	0.0016	42.4	0.0180	0.2901	0.0159
+MI; 5	0.0016	43.8	0.0189	0.3056	0.0158
+MI; 10	0.0016	43.2	0.0189	0.3049	0.0159
+MI-B; 5	0.0019	97.6	0.0673	1.7403	0.0106
+MI-B; 10	0.0019	98.8	0.0656	1.8367	0.0095
X=3					
HMM	-0.0090	82.6	0.0427	1.0350	0.0138
+SI	-0.0089	32.8	0.0122	0.2921	0.0139
+MI; 5	-0.0089	35.4	0.0138	0.3364	0.0138
+MI; 10	-0.0089	34.6	0.0136	0.3315	0.0138
+MI-B; 5	-0.0114	83.8	0.0440	1.4060	0.0141
+MI-B; 10	-0.0116	87.0	0.0433	1.4328	0.0141

Table C5: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		-0.0003	95.4	0.0638	1.7204	0.0095
+SI	Conditional	-0.0004	77.0	0.0243	0.6239	0.0099
	Marginal	-0.0003	79.2	0.0243	0.6211	0.0100
+MI; 5	Conditional	-0.0004	86.0	0.0286	0.7633	0.0096
	Marginal	-0.0003	84.6	0.0270	0.7212	0.0096
+MI; 10	Conditional	-0.0004	86.8	0.0284	0.7567	0.0096
	Marginal	-0.0004	85.6	0.0267	0.7160	0.0095
+MI-B; 5	Conditional	0.0063	89.4	0.0406	1.0892	0.0116
	Marginal	0.0064	89.2	0.0398	1.0691	0.0116
+MI-B; 10	Conditional	0.0064	91.0	0.0397	1.0806	0.0114
	Marginal	0.0064	89.8	0.0386	1.0472	0.0114

Table C6: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		0.0001	94.2	0.0246	1.3947	0.0045
+SI	Conditional	-0.0003	93.2	0.0243	0.9210	0.0047
	Marginal	-0.0002	92.6	0.0243	0.9032	0.0047
+MI; 5	Conditional	-0.0002	94.4	0.0248	0.9501	0.0045
	Marginal	-0.0002	94.2	0.0248	0.9357	0.0045
+MI; 10	Conditional	-0.0002	94.4	0.0248	0.9452	0.0045
	Marginal	-0.0002	94.2	0.0247	0.9387	0.0045
+MI-B; 5	Conditional	0.0005	90.0	0.0174	0.9591	0.0047
	Marginal	0.0006	91.4	0.0170	0.9522	0.0047
+MI-B; 10	Conditional	0.0006	92.8	0.0171	0.9620	0.0046
	Marginal	0.0006	90.8	0.0165	0.9274	0.0046

Table C7: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		-0.0002	94.6	0.0256	0.9851	0.0066
+SI	Conditional	-0.0003	93.2	0.0243	0.9210	0.0067
	Marginal	-0.0002	92.6	0.0243	0.9032	0.0069
+MI; 5	Conditional	-0.0002	94.4	0.0248	0.9501	0.0067
	Marginal	-0.0002	94.2	0.0248	0.9357	0.0067
+MI; 10	Conditional	-0.0002	94.4	0.0248	0.9452	0.0067
	Marginal	-0.0002	94.2	0.0247	0.9387	0.0067
+MI-B; 5	Conditional	0.0026	92.8	0.0260	0.9810	0.0073
	Marginal	0.0026	92.8	0.0260	0.9853	0.0072
+MI-B; 10	Conditional	0.0026	93.2	0.0259	0.9797	0.0072
	Marginal	0.0026	92.8	0.0259	0.9780	0.0072

Table C8: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		0.0001	94.8	0.0114	0.9653	0.0030
+SI	Conditional	0.0001	92.8	0.0109	0.8993	0.0031
	Marginal	0.0001	93.2	0.0109	0.8906	0.0031
+MI; 5	Conditional	0.0001	94.4	0.0111	0.9341	0.0030
	Marginal	0.0001	94.2	0.0111	0.9258	0.0030
+MI; 10	Conditional	0.0001	94.2	0.0111	0.9311	0.0030
	Marginal	0.0001	94.2	0.0110	0.9260	0.0030
+MI-B; 5	Conditional	0.0010	93.0	0.0114	0.9395	0.0032
	Marginal	0.0010	92.8	0.0113	0.9352	0.0033
+MI-B; 10	Conditional	0.0010	93.2	0.0113	0.9463	0.0032
	Marginal	0.0010	94.0	0.0113	0.9380	0.0032

Table C9: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0020	95.2	0.0439	1.0949	0.0104
+SI	Conditional	-0.0019	81.6	0.0288	0.6915	0.0108
	Marginal	-0.0019	77.2	0.0266	0.6295	0.0110
+MI; 5	Conditional	-0.0019	88.6	0.0338	0.8388	0.0105
	Marginal	-0.0019	85.2	0.0296	0.7275	0.0106
+MI; 10	Conditional	-0.0019	89.4	0.0335	0.8342	0.0104
	Marginal	-0.0019	84.0	0.0294	0.7258	0.0105
+MI-B; 5	Conditional	0.0038	93.0	0.0424	1.0427	0.0112
	Marginal	0.0040	89.4	0.0399	1.0013	0.0111
+MI-B; 10	Conditional	0.0038	92.0	0.0421	1.0729	0.0108
	Marginal	0.0038	90.8	0.0391	0.9925	0.0109
$Q_1 = 2$						
HMM		0.0028	96.0	0.1446	2.3169	0.0162
+SI	Conditional	0.0026	82.8	0.0487	0.7183	0.0175
	Marginal	0.0023	81.4	0.0455	0.6855	0.0171
+MI; 5	Conditional	0.0025	89.8	0.0563	0.8818	0.0165
	Marginal	0.0024	88.0	0.0505	0.8035	0.0162
+MI; 10	Conditional	0.0026	90.6	0.0561	0.8839	0.0164
	Marginal	0.0024	88.2	0.0501	0.7979	0.0162
+MI-B; 5	Conditional	-0.0078	90.6	0.0699	1.0927	0.0183
	Marginal	-0.0079	90.4	0.0648	1.0081	0.0184
+MI-B; 10	Conditional	-0.0078	91.4	0.0678	1.0667	0.0181
	Marginal	-0.0077	89.6	0.0637	1.0204	0.0178
$Q_1 = 3$						
HMM		0.0004	95.8	0.1366	1.0806	0.0322
+SI	Conditional	0.0000	83.8	0.1038	0.7326	0.0362
	Marginal	0.0013	83.2	0.1000	0.7275	0.0351
+MI; 5	Conditional	-0.0002	94.0	0.1231	0.9632	0.0328
	Marginal	0.0008	91.2	0.1141	0.8853	0.0330
+MI; 10	Conditional	0.0001	94.2	0.1216	0.9568	0.0325
	Marginal	0.0009	91.4	0.1130	0.8902	0.0324
+MI-B; 5	Conditional	0.0004	95.6	0.1500	1.0535	0.0369
	Marginal	0.0003	94.4	0.1443	1.0305	0.0362
+MI-B; 10	Conditional	0.0010	94.4	0.1473	1.0602	0.0357
	Marginal	0.0006	94.0	0.1411	1.0130	0.0358

Table C10: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0007	94.8	0.0293	0.9817	0.0077
+SI	Conditional	-0.0006	93.2	0.0290	0.9570	0.0054
	Marginal	-0.0005	93.2	0.0289	0.9575	0.0052
+MI; 5	Conditional	-0.0006	94.0	0.0295	0.9790	0.0050
	Marginal	-0.0005	93.8	0.0294	0.9728	0.0050
+MI; 10	Conditional	-0.0006	94.0	0.0295	0.9772	0.0050
	Marginal	-0.0006	93.6	0.0293	0.9735	0.0050
+MI-B; 5	Conditional	-0.0003	91.4	0.0189	0.9475	0.0052
	Marginal	-0.0003	90.6	0.0176	0.8911	0.0051
+MI-B; 10	Conditional	-0.0002	92.4	0.0188	0.9579	0.0051
	Marginal	-0.0002	90.6	0.0174	0.8859	0.0050
$Q_1 = 2$						
HMM		0.0008	94.4	0.0493	0.9808	0.0128
+SI	Conditional	0.0004	94.2	0.0489	0.9455	0.0077
	Marginal	0.0003	92.8	0.0488	0.9384	0.0078
+MI; 5	Conditional	0.0004	93.8	0.0497	0.9753	0.0076
	Marginal	0.0004	94.4	0.0495	0.9709	0.0076
+MI; 10	Conditional	0.0004	93.6	0.0496	0.9741	0.0076
	Marginal	0.0004	94.0	0.0494	0.9729	0.0075
+MI-B; 5	Conditional	-0.0004	94.8	0.0303	1.0142	0.0077
	Marginal	-0.0004	94.0	0.0282	0.9630	0.0076
+MI-B; 10	Conditional	-0.0004	95.0	0.0297	1.0140	0.0075
	Marginal	-0.0004	93.2	0.0275	0.9399	0.0075
$Q_1 = 3$						
HMM		-0.0003	94.8	0.1072	1.0065	0.0272
+SI	Conditional	-0.0001	93.6	0.1042	0.9465	0.0167
	Marginal	-0.0003	94.0	0.1041	0.9408	0.0167
+MI; 5	Conditional	0.0001	95.0	0.1073	0.9936	0.0156
	Marginal	-0.0000	94.8	0.1071	0.9955	0.0156
+MI; 10	Conditional	0.0002	94.6	0.1071	0.9963	0.0154
	Marginal	0.0000	95.4	0.1069	0.9961	0.0154
+MI-B; 5	Conditional	0.0008	91.2	0.0613	0.9466	0.0167
	Marginal	0.0006	92.2	0.0581	0.9085	0.0165
+MI-B; 10	Conditional	0.0007	92.4	0.0601	0.9546	0.0161
	Marginal	0.0007	92.6	0.0575	0.9242	0.0159

Table C11: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0007	94.8	0.0293	0.9821	0.0077
+SI	Conditional	-0.0006	93.2	0.0290	0.9570	0.0078
	Marginal	-0.0005	93.2	0.0289	0.9575	0.0077
+MI; 5	Conditional	-0.0006	94.0	0.0295	0.9790	0.0077
	Marginal	-0.0005	93.8	0.0294	0.9728	0.0077
+MI; 10	Conditional	-0.0006	94.0	0.0295	0.9772	0.0077
	Marginal	-0.0006	93.6	0.0293	0.9735	0.0077
+MI-B; 5	Conditional	0.0039	89.8	0.0299	0.9978	0.0086
	Marginal	0.0040	89.4	0.0298	0.9978	0.0086
+MI-B; 10	Conditional	0.0039	89.6	0.0298	0.9910	0.0086
	Marginal	0.0040	89.4	0.0297	0.9860	0.0086
$Q_1 = 2$						
HMM		0.0009	94.4	0.0493	0.9787	0.0129
+SI	Conditional	0.0004	94.2	0.0489	0.9455	0.0132
	Marginal	0.0003	92.8	0.0488	0.9384	0.0133
+MI; 5	Conditional	0.0004	93.8	0.0497	0.9753	0.0130
	Marginal	0.0004	94.4	0.0495	0.9709	0.0130
+MI; 10	Conditional	0.0004	93.6	0.0496	0.9741	0.0130
	Marginal	0.0004	94.0	0.0494	0.9729	0.0130
+MI-B; 5	Conditional	-0.0064	90.4	0.0505	0.9836	0.0146
	Marginal	-0.0065	90.0	0.0504	0.9807	0.0146
+MI-B; 10	Conditional	-0.0065	90.0	0.0503	0.9852	0.0146
	Marginal	-0.0065	89.4	0.0502	0.9816	0.0146
$Q_1 = 3$						
HMM		-0.0004	94.8	0.1072	1.0058	0.0272
+SI	Conditional	-0.0001	93.6	0.1042	0.9465	0.0281
	Marginal	-0.0003	94.0	0.1041	0.9408	0.0282
+MI; 5	Conditional	0.0001	95.0	0.1073	0.9936	0.0275
	Marginal	-0.0000	94.8	0.1071	0.9955	0.0274
+MI; 10	Conditional	0.0002	94.6	0.1071	0.9963	0.0274
	Marginal	0.0000	95.4	0.1069	0.9961	0.0274
+MI-B; 5	Conditional	0.0002	94.8	0.1120	0.9813	0.0291
	Marginal	0.0003	94.6	0.1120	0.9931	0.0288
+MI-B; 10	Conditional	0.0001	95.2	0.1114	0.9907	0.0287
	Marginal	0.0000	94.6	0.1112	0.9831	0.0289

Table C12: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		0.0000	94.4	0.0131	0.9511	0.0035
+SI	Conditional	0.0001	93.0	0.0129	0.9204	0.0036
	Marginal	0.0000	93.2	0.0129	0.9060	0.0036
+MI; 5	Conditional	0.0001	93.6	0.0132	0.9474	0.0035
	Marginal	0.0000	93.8	0.0131	0.9403	0.0036
+MI; 10	Conditional	0.0001	94.2	0.0131	0.9483	0.0035
	Marginal	0.0000	93.4	0.0131	0.9418	0.0035
+MI-B; 5	Conditional	0.0011	93.8	0.0133	0.9388	0.0038
	Marginal	0.0011	93.8	0.0132	0.9391	0.0038
+MI-B; 10	Conditional	0.0011	93.2	0.0133	0.9450	0.0037
	Marginal	0.0011	93.8	0.0132	0.9388	0.0037
$Q_1 = 2$						
HMM		0.0005	94.2	0.0221	1.0269	0.0055
+SI	Conditional	0.0003	93.6	0.0219	0.9832	0.0057
	Marginal	0.0003	94.6	0.0219	0.9873	0.0057
+MI; 5	Conditional	0.0003	94.4	0.0223	1.0214	0.0056
	Marginal	0.0003	94.4	0.0222	1.0106	0.0056
+MI; 10	Conditional	0.0003	94.0	0.0222	1.0199	0.0056
	Marginal	0.0003	94.4	0.0221	1.0099	0.0056
+MI-B; 5	Conditional	-0.0015	93.8	0.0225	1.0106	0.0059
	Marginal	-0.0014	93.8	0.0224	1.0082	0.0058
+MI-B; 10	Conditional	-0.0015	93.6	0.0224	1.0139	0.0058
	Marginal	-0.0015	93.8	0.0224	1.0109	0.0058
$Q_1 = 3$						
HMM		-0.0000	94.6	0.0479	1.0154	0.0120
+SI	Conditional	0.0000	93.8	0.0466	0.9636	0.0123
	Marginal	0.0002	94.0	0.0466	0.9584	0.0124
+MI; 5	Conditional	0.0001	94.6	0.0481	1.0061	0.0122
	Marginal	0.0002	94.4	0.0479	1.0055	0.0121
+MI; 10	Conditional	0.0002	94.8	0.0479	1.0014	0.0122
	Marginal	0.0002	94.6	0.0478	1.0020	0.0122
+MI-B; 5	Conditional	0.0003	95.0	0.0488	0.9929	0.0125
	Marginal	0.0003	95.2	0.0488	1.0017	0.0124
+MI-B; 10	Conditional	0.0003	94.8	0.0486	1.0008	0.0124
	Marginal	0.0003	95.0	0.0486	0.9988	0.0124

Table C13: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_2 = 1$						
HMM		0.0010	93.0	0.0811	1.4499	0.0143
+SI	Conditional	0.0010	82.2	0.0421	0.7105	0.0151
	Marginal	0.0014	84.4	0.0421	0.7260	0.0148
+MI; 5	Conditional	0.0010	91.0	0.0496	0.8896	0.0143
	Marginal	0.0012	89.4	0.0468	0.8409	0.0143
+MI; 10	Conditional	0.0011	91.6	0.0491	0.8837	0.0142
	Marginal	0.0011	89.8	0.0465	0.8418	0.0141
+MI-B; 5	Conditional	0.0081	91.0	0.0602	1.0624	0.0167
	Marginal	0.0081	91.0	0.0591	1.0326	0.0168
+MI-B; 10	Conditional	0.0080	92.4	0.0599	1.0684	0.0165
	Marginal	0.0080	91.6	0.0578	1.0266	0.0165
$Q_2 = 2$						
HMM		-0.0003	94.0	0.0789	1.4152	0.0142
+SI	Conditional	-0.0003	84.0	0.0421	0.7204	0.0149
	Marginal	-0.0007	83.8	0.0421	0.7222	0.0149
+MI; 5	Conditional	-0.0006	91.4	0.0492	0.8828	0.0143
	Marginal	-0.0006	89.4	0.0465	0.8366	0.0142
+MI; 10	Conditional	-0.0005	91.2	0.0491	0.8826	0.0142
	Marginal	-0.0006	88.6	0.0463	0.8334	0.0142
+MI-B; 5	Conditional	0.0061	93.2	0.0611	1.0943	0.0157
	Marginal	0.0063	90.2	0.0591	1.0576	0.0157
+MI-B; 10	Conditional	0.0062	92.4	0.0597	1.0694	0.0156
	Marginal	0.0062	92.2	0.0578	1.0341	0.0156
$Q_2 = 3$						
HMM		-0.0018	91.8	0.0800	1.3261	0.0155
+SI	Conditional	-0.0018	81.6	0.0421	0.6517	0.0166
	Marginal	-0.0017	80.0	0.0421	0.6642	0.0162
+MI; 5	Conditional	-0.0017	89.6	0.0493	0.8278	0.0154
	Marginal	-0.0016	87.4	0.0466	0.7778	0.0154
+MI; 10	Conditional	-0.0016	88.8	0.0492	0.8252	0.0153
	Marginal	-0.0016	87.2	0.0463	0.7802	0.0152
+MI-B; 5	Conditional	0.0048	93.4	0.0615	1.0297	0.0162
	Marginal	0.0049	92.6	0.0589	0.9895	0.0161
+MI-B; 10	Conditional	0.0051	93.6	0.0600	1.0202	0.0159
	Marginal	0.0051	91.8	0.0576	0.9747	0.0160

Table C14: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.80 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_2 = 1$						
HMM		-0.0000	91.6	0.0391	0.8797	0.0113
+SI	Conditional	0.0002	93.6	0.0421	0.9483	0.0072
	Marginal	0.0002	94.0	0.0421	0.9463	0.0069
+MI; 5	Conditional	0.0003	95.4	0.0430	0.9807	0.0066
	Marginal	0.0003	94.8	0.0429	0.9665	0.0066
+MI; 10	Conditional	0.0003	95.0	0.0429	0.9773	0.0065
	Marginal	0.0003	95.0	0.0428	0.9687	0.0066
+MI-B; 5	Conditional	0.0005	92.8	0.0260	0.9870	0.0068
	Marginal	0.0005	93.4	0.0252	0.9747	0.0067
+MI-B; 10	Conditional	0.0005	94.6	0.0259	1.0037	0.0066
	Marginal	0.0005	94.2	0.0247	0.9565	0.0066
$Q_2 = 2$						
HMM		-0.0008	94.0	0.0391	0.9135	0.0109
+SI	Conditional	-0.0012	93.8	0.0421	0.9550	0.0068
	Marginal	-0.0010	93.0	0.0421	0.9490	0.0069
+MI; 5	Conditional	-0.0011	93.8	0.0430	0.9796	0.0066
	Marginal	-0.0010	93.6	0.0429	0.9787	0.0067
+MI; 10	Conditional	-0.0011	93.8	0.0429	0.9758	0.0066
	Marginal	-0.0011	93.4	0.0428	0.9796	0.0066
+MI-B; 5	Conditional	0.0005	94.4	0.0263	1.0138	0.0067
	Marginal	0.0005	92.8	0.0249	0.9603	0.0067
+MI-B; 10	Conditional	0.0005	93.8	0.0257	0.9882	0.0067
	Marginal	0.0005	93.0	0.0248	0.9578	0.0067
$Q_2 = 3$						
HMM		0.0001	92.8	0.0392	0.8988	0.0111
+SI	Conditional	0.0002	92.4	0.0421	0.9320	0.0074
	Marginal	0.0001	94.0	0.0421	0.9330	0.0073
+MI; 5	Conditional	0.0001	94.2	0.0430	0.9618	0.0071
	Marginal	0.0001	95.0	0.0428	0.9605	0.0071
+MI; 10	Conditional	0.0001	93.8	0.0429	0.9642	0.0070
	Marginal	0.0001	94.6	0.0428	0.9613	0.0070
+MI-B; 5	Conditional	0.0007	91.0	0.0261	0.9173	0.0074
	Marginal	0.0007	91.4	0.0250	0.8978	0.0072
+MI-B; 10	Conditional	0.0008	92.6	0.0257	0.9239	0.0072
	Marginal	0.0008	91.4	0.0245	0.8840	0.0071

Table C15: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_2 = 1$						
HMM		0.0000	91.4	0.0391	0.8776	0.0113
+SI	Conditional	0.0002	93.6	0.0421	0.9483	0.0113
	Marginal	0.0002	94.0	0.0421	0.9463	0.0113
+MI; 5	Conditional	0.0003	95.4	0.0430	0.9807	0.0112
	Marginal	0.0003	94.8	0.0429	0.9665	0.0113
+MI; 10	Conditional	0.0003	95.0	0.0429	0.9773	0.0112
	Marginal	0.0003	95.0	0.0428	0.9687	0.0113
+MI-B; 5	Conditional	0.0032	95.0	0.0446	1.0018	0.0118
	Marginal	0.0032	94.0	0.0446	0.9988	0.0118
+MI-B; 10	Conditional	0.0031	94.6	0.0445	0.9970	0.0118
	Marginal	0.0031	94.2	0.0444	0.9949	0.0118
$Q_2 = 2$						
HMM		-0.0008	94.0	0.0391	0.9143	0.0109
+SI	Conditional	-0.0012	93.8	0.0421	0.9550	0.0113
	Marginal	-0.0010	93.0	0.0421	0.9490	0.0114
+MI; 5	Conditional	-0.0011	93.8	0.0430	0.9796	0.0112
	Marginal	-0.0010	93.6	0.0429	0.9787	0.0112
+MI; 10	Conditional	-0.0011	93.8	0.0429	0.9758	0.0113
	Marginal	-0.0011	93.4	0.0428	0.9796	0.0112
+MI-B; 5	Conditional	0.0018	95.0	0.0446	1.0019	0.0115
	Marginal	0.0018	95.0	0.0446	1.0054	0.0115
+MI-B; 10	Conditional	0.0017	95.0	0.0445	1.0000	0.0115
	Marginal	0.0017	95.0	0.0444	1.0034	0.0114
$Q_2 = 3$						
HMM		0.0002	92.8	0.0392	0.8995	0.0111
+SI	Conditional	0.0002	92.4	0.0421	0.9320	0.0115
	Marginal	0.0001	94.0	0.0421	0.9330	0.0115
+MI; 5	Conditional	0.0001	94.2	0.0430	0.9618	0.0114
	Marginal	0.0001	95.0	0.0428	0.9605	0.0114
+MI; 10	Conditional	0.0001	93.8	0.0429	0.9642	0.0113
	Marginal	0.0001	94.6	0.0428	0.9613	0.0113
+MI-B; 5	Conditional	0.0029	93.6	0.0446	0.9887	0.0119
	Marginal	0.0029	94.2	0.0446	0.9904	0.0119
+MI-B; 10	Conditional	0.0029	93.6	0.0444	0.9908	0.0118
	Marginal	0.0029	93.4	0.0444	0.9852	0.0119

Table C16: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_2 = 1$						
HMM		-0.0000	92.0	0.0176	0.9322	0.0050
+SI	Conditional	-0.0000	93.4	0.0188	0.9175	0.0051
	Marginal	0.0000	92.6	0.0188	0.9172	0.0051
+MI; 5	Conditional	-0.0000	94.6	0.0192	0.9506	0.0050
	Marginal	-0.0000	95.4	0.0192	0.9470	0.0050
+MI; 10	Conditional	-0.0000	95.2	0.0192	0.9503	0.0050
	Marginal	-0.0000	95.2	0.0191	0.9474	0.0050
+MI-B; 5	Conditional	0.0009	94.0	0.0195	0.9588	0.0052
	Marginal	0.0009	93.8	0.0195	0.9469	0.0052
+MI-B; 10	Conditional	0.0009	93.6	0.0194	0.9511	0.0052
	Marginal	0.0009	93.8	0.0194	0.9523	0.0052
$Q_2 = 2$						
HMM		0.0002	93.8	0.0176	0.9212	0.0049
+SI	Conditional	0.0002	94.4	0.0188	0.9505	0.0051
	Marginal	0.0001	94.6	0.0188	0.9627	0.0050
+MI; 5	Conditional	0.0002	95.8	0.0192	0.9917	0.0049
	Marginal	0.0001	95.2	0.0192	0.9903	0.0049
+MI; 10	Conditional	0.0002	95.6	0.0192	0.9937	0.0049
	Marginal	0.0001	94.8	0.0191	0.9908	0.0049
+MI-B; 5	Conditional	0.0010	95.0	0.0195	1.0030	0.0051
	Marginal	0.0011	95.0	0.0195	1.0068	0.0050
+MI-B; 10	Conditional	0.0010	95.0	0.0195	1.0057	0.0050
	Marginal	0.0010	94.4	0.0194	1.0052	0.0050
$Q_2 = 3$						
HMM		-0.0000	92.0	0.0176	0.8907	0.0050
+SI	Conditional	-0.0000	93.4	0.0188	0.9175	0.0052
	Marginal	0.0000	92.6	0.0188	0.9172	0.0052
+MI; 5	Conditional	-0.0000	94.6	0.0192	0.9506	0.0052
	Marginal	-0.0000	95.4	0.0192	0.9470	0.0052
+MI; 10	Conditional	-0.0000	95.2	0.0192	0.9503	0.0051
	Marginal	-0.0000	95.2	0.0191	0.9474	0.0051
+MI-B; 5	Conditional	0.0009	94.0	0.0195	0.9588	0.0053
	Marginal	0.0009	93.8	0.0195	0.9469	0.0053
+MI-B; 10	Conditional	0.0009	93.6	0.0194	0.9511	0.0053
	Marginal	0.0009	93.8	0.0194	0.9523	0.0053

Table C17: Results for X in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators and missing data with a MCAR structure

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0086	86.2	0.0667	1.1305	0.0173
+SI	0.0087	43.4	0.0193	0.3250	0.0174
+MI; 5	0.0086	45.8	0.0204	0.3467	0.0173
+MI; 10	0.0086	45.2	0.0203	0.3438	0.0173
+MI-B; 5	0.0112	92.6	0.0668	1.7349	0.0152
+MI-B; 10	0.0108	96.6	0.0649	1.8440	0.0142
X=2					
HMM	0.0018	92.4	0.0675	1.0550	0.0164
+SI	0.0017	38.4	0.0180	0.2819	0.0164
+MI; 5	0.0017	40.4	0.0190	0.2975	0.0164
+MI; 10	0.0017	41.0	0.0189	0.2960	0.0164
+MI-B; 5	0.0015	96.6	0.0675	1.7386	0.0106
+MI-B; 10	0.0019	98.4	0.0670	1.8490	0.0097
X=3					
HMM	-0.0104	81.2	0.0468	1.0219	0.0156
+SI	-0.0103	30.4	0.0121	0.2612	0.0157
+MI; 5	-0.0104	34.8	0.0137	0.2987	0.0157
+MI; 10	-0.0104	33.0	0.0136	0.2971	0.0156
+MI-B; 5	-0.0127	82.0	0.0490	1.3588	0.0160
+MI-B; 10	-0.0127	87.0	0.0480	1.4286	0.0155

Table C18: Results for X in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $2,000$ and probabilities of correct classification of 0.80 for both indicators and missing data with a MAR structure

	bias	% cov	CI width	se/sd	RMSE
X=1					
HMM	0.0093	86.6	0.0689	1.1932	0.0174
+SI	0.0094	43.2	0.0193	0.3313	0.0176
+MI; 5	0.0094	45.8	0.0204	0.3518	0.0175
+MI; 10	0.0094	45.8	0.0204	0.3518	0.0175
+MI-B; 5	0.0109	93.0	0.0683	1.8334	0.0148
+MI-B; 10	0.0111	96.0	0.0665	1.9856	0.0142
X=2					
HMM	0.0011	93.2	0.0723	1.1759	0.0157
+SI	0.0011	39.4	0.0180	0.2934	0.0157
+MI; 5	0.0011	41.6	0.0190	0.3087	0.0157
+MI; 10	0.0011	41.6	0.0190	0.3087	0.0157
+MI-B; 5	0.0012	96.6	0.0672	1.7671	0.0103
+MI-B; 10	0.0012	99.6	0.0669	1.9676	0.0090
X=3					
HMM	-0.0105	79.6	0.0471	1.0151	0.0158
+SI	-0.0105	28.8	0.0121	0.2563	0.0160
+MI; 5	-0.0105	32.8	0.0137	0.2932	0.0159
+MI; 10	-0.0105	32.8	0.0137	0.2932	0.0159
+MI-B; 5	-0.0121	80.8	0.0481	1.3993	0.0153
+MI-B; 10	-0.0123	85.0	0.0479	1.4979	0.0149

Table C19: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MCAR structure. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		0.0001	94.4	0.0127	0.9769	0.0033
+SI	Conditional	0.0001	89.6	0.0109	0.8037	0.0034
	Marginal	0.0001	88.4	0.0109	0.8139	0.0034
+MI; 5	Conditional	0.0001	92.6	0.0114	0.8661	0.0034
	Marginal	0.0001	91.4	0.0113	0.8630	0.0033
+MI; 10	Conditional	0.0001	92.0	0.0114	0.8717	0.0033
	Marginal	0.0001	91.8	0.0113	0.8602	0.0034
+MI-B; 5	Conditional	0.0005	92.4	0.0125	0.9316	0.0035
	Marginal	0.0005	93.0	0.0123	0.9274	0.0034
+MI-B; 10	Conditional	0.0005	93.4	0.0124	0.9488	0.0034
	Marginal	0.0005	93.6	0.0123	0.9441	0.0034

Table C20: Results for **L** in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MAR structure. Note that as **L** has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
HMM		0.0001	94.8	0.0127	0.9662	0.0033
+SI	Conditional	0.0002	89.8	0.0109	0.8023	0.0035
	Marginal	0.0002	88.4	0.0109	0.8083	0.0034
+MI; 5	Conditional	0.0001	91.0	0.0115	0.8660	0.0034
	Marginal	0.0002	91.2	0.0113	0.8610	0.0034
+MI; 10	Conditional	0.0001	91.0	0.0115	0.8660	0.0034
	Marginal	0.0002	91.2	0.0113	0.8610	0.0034
+MI-B; 5	Conditional	0.0005	93.0	0.0125	0.9337	0.0035
	Marginal	0.0006	93.2	0.0124	0.9300	0.0035
+MI-B; 10	Conditional	0.0006	94.0	0.0124	0.9327	0.0034
	Marginal	0.0006	93.2	0.0123	0.9273	0.0034

Table C21: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MCAR structure. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0001	94.0	0.0144	0.9528	0.0039
+SI	Conditional	-0.0001	89.2	0.0129	0.8198	0.0040
	Marginal	-0.0001	89.0	0.0128	0.8239	0.0040
+MI; 5	Conditional	-0.0001	91.4	0.0136	0.8898	0.0039
	Marginal	-0.0001	91.6	0.0133	0.8724	0.0039
+MI; 10	Conditional	-0.0001	91.8	0.0136	0.8929	0.0039
	Marginal	-0.0001	91.4	0.0133	0.8743	0.0039
+MI-B; 5	Conditional	0.0005	93.0	0.0144	0.9320	0.0040
	Marginal	0.0005	91.6	0.0141	0.9140	0.0040
+MI-B; 10	Conditional	0.0005	92.4	0.0144	0.9436	0.0039
	Marginal	0.0005	91.8	0.0141	0.9251	0.0039
$Q_1 = 2$						
HMM		0.0006	95.2	0.0240	1.0386	0.0059
+SI	Conditional	0.0004	91.8	0.0219	0.8963	0.0062
	Marginal	0.0004	93.2	0.0217	0.9061	0.0061
+MI; 5	Conditional	0.0004	93.2	0.0230	0.9687	0.0061
	Marginal	0.0004	94.0	0.0225	0.9473	0.0061
+MI; 10	Conditional	0.0004	94.2	0.0229	0.9761	0.0060
	Marginal	0.0004	93.8	0.0224	0.9514	0.0060
+MI-B; 5	Conditional	-0.0008	94.4	0.0241	1.0214	0.0061
	Marginal	-0.0008	93.6	0.0237	1.0034	0.0061
+MI-B; 10	Conditional	-0.0008	94.8	0.0240	1.0192	0.0060
	Marginal	-0.0008	94.6	0.0235	0.9971	0.0061
$Q_1 = 3$						
HMM		0.0002	95.2	0.0495	1.0208	0.0124
+SI	Conditional	0.0001	92.2	0.0466	0.9132	0.0130
	Marginal	0.0003	94.0	0.0464	0.9172	0.0129
+MI; 5	Conditional	0.0002	95.0	0.0493	0.9932	0.0127
	Marginal	0.0001	95.0	0.0487	0.9921	0.0125
+MI; 10	Conditional	0.0003	95.0	0.0490	0.9931	0.0126
	Marginal	0.0002	94.4	0.0485	0.9882	0.0131
+MI-B; 5	Conditional	0.0004	95.2	0.0505	1.0039	0.0128
	Marginal	0.0003	94.2	0.0498	0.9942	0.0128
+MI-B; 10	Conditional	0.0003	94.8	0.0502	1.0053	0.0127
	Marginal	0.0004	95.0	0.0497	0.9921	0.0128

Table C22: Results for $\mathbf{L} \mid Q_1$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MAR structure. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0001	92.6	0.0144	0.9510	0.0039
+SI	Conditional	0.0000	89.4	0.0129	0.8425	0.0039
	Marginal	-0.0000	87.4	0.0128	0.8200	0.0040
+MI; 5	Conditional	-0.0000	91.4	0.0136	0.8935	0.0039
	Marginal	-0.0000	90.0	0.0132	0.8755	0.0039
+MI; 10	Conditional	-0.0000	91.4	0.0136	0.8935	0.0039
	Marginal	-0.0000	90.0	0.0132	0.8755	0.0039
+MI-B; 5	Conditional	0.0006	90.8	0.0144	0.9235	0.0040
	Marginal	0.0005	90.4	0.0141	0.9098	0.0040
+MI-B; 10	Conditional	0.0006	92.0	0.0143	0.9303	0.0040
	Marginal	0.0006	91.4	0.0140	0.9128	0.0040
$Q_1 = 2$						
HMM		0.0006	94.6	0.0240	1.0425	0.0059
+SI	Conditional	0.0005	91.4	0.0219	0.8981	0.0062
	Marginal	0.0004	91.2	0.0217	0.9110	0.0061
+MI; 5	Conditional	0.0004	94.0	0.0230	0.9822	0.0060
	Marginal	0.0004	93.8	0.0225	0.9664	0.0060
+MI; 10	Conditional	0.0004	94.0	0.0230	0.9822	0.0060
	Marginal	0.0004	93.8	0.0225	0.9664	0.0060
+MI-B; 5	Conditional	-0.0008	95.2	0.0242	1.0314	0.0060
	Marginal	-0.0007	93.8	0.0237	1.0132	0.0060
+MI-B; 10	Conditional	-0.0007	94.4	0.0240	1.0244	0.0060
	Marginal	-0.0007	94.6	0.0237	1.0146	0.0060
$Q_1 = 3$						
HMM		0.0002	94.4	0.0495	1.0003	0.0126
+SI	Conditional	0.0007	92.2	0.0466	0.8984	0.0132
	Marginal	0.0007	92.2	0.0464	0.8981	0.0132
+MI; 5	Conditional	0.0005	94.4	0.0493	0.9838	0.0128
	Marginal	0.0004	94.2	0.0487	0.9660	0.0129
+MI; 10	Conditional	0.0005	94.4	0.0493	0.9838	0.0128
	Marginal	0.0004	94.2	0.0487	0.9660	0.0129
+MI-B; 5	Conditional	0.0004	95.0	0.0503	0.9961	0.0129
	Marginal	0.0005	95.0	0.0502	1.0005	0.0128
+MI-B; 10	Conditional	0.0004	95.2	0.0503	0.9917	0.0129
	Marginal	0.0005	94.8	0.0498	0.9879	0.0129

Table C23: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MCAR structure. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		0.0003	95.2	0.0189	0.9449	0.0051
+SI	Conditional	0.0003	91.2	0.0188	0.8813	0.0055
	Marginal	0.0003	92.4	0.0188	0.8948	0.0054
+MI; 5	Conditional	0.0003	93.8	0.0198	0.9500	0.0053
	Marginal	0.0003	94.6	0.0196	0.9491	0.0053
+MI; 10	Conditional	0.0003	94.0	0.0198	0.9558	0.0053
	Marginal	0.0003	94.4	0.0196	0.9490	0.0053
+MI-B; 5	Conditional	0.0007	94.6	0.0206	0.9903	0.0054
	Marginal	0.0007	94.6	0.0204	0.9783	0.0054
+MI-B; 10	Conditional	0.0007	94.8	0.0205	0.9897	0.0053
	Marginal	0.0007	94.0	0.0203	0.9832	0.0053
$Q_1 = 2$						
HMM		0.0001	94.8	0.0189	0.9444	0.0051
+SI	Conditional	0.0001	92.8	0.0188	0.9135	0.0053
	Marginal	0.0001	91.8	0.0188	0.9079	0.0053
+MI; 5	Conditional	0.0001	95.0	0.0199	0.9961	0.0051
	Marginal	0.0001	95.0	0.0197	0.9793	0.0051
+MI; 10	Conditional	0.0001	95.0	0.0198	0.9962	0.0051
	Marginal	0.0002	95.2	0.0196	0.9769	0.0051
+MI-B; 5	Conditional	0.0006	95.8	0.0205	1.0081	0.0052
	Marginal	0.0005	95.8	0.0203	1.0042	0.0052
+MI-B; 10	Conditional	0.0006	95.6	0.0205	1.0194	0.0052
	Marginal	0.0005	96.0	0.0203	1.0095	0.0052
$Q_1 = 3$						
HMM		-0.0002	92.4	0.0189	0.8928	0.0054
+SI	Conditional	-0.0002	90.0	0.0188	0.8489	0.0057
	Marginal	-0.0002	91.6	0.0188	0.8576	0.0056
+MI; 5	Conditional	-0.0002	94.2	0.0198	0.9139	0.0055
	Marginal	-0.0002	94.0	0.0196	0.9065	0.0055
+MI; 10	Conditional	-0.0001	94.2	0.0198	0.9141	0.0055
	Marginal	-0.0002	93.8	0.0196	0.9037	0.0055
+MI-B; 5	Conditional	0.0003	93.8	0.0207	0.9480	0.0056
	Marginal	0.0003	94.0	0.0204	0.9326	0.0056
+MI-B; 10	Conditional	0.0003	94.4	0.0205	0.9482	0.0055
	Marginal	0.0003	94.4	0.0203	0.9442	0.0055

Table C24: Results for $\mathbf{L} \mid Q_2$ in terms of five different performance measures and obtained after using six different strategies and two different imputation procedures to apply the longitudinal extension of the MILC method in the rows. Results obtained for the simulation condition with a sample size of $10,000$ and probabilities of correct classification of 0.95 for both indicators and missing data with a MAR structure. Note that as $\mathbf{L}$ has only two categories, the results are symmetrical and therefore only the results of $\mathbf{L} = 1$ are shown

		bias	% cov	CI width	se/sd	RMSE
$Q_1 = 1$						
HMM		-0.0001	94.8	0.0189	0.9627	0.0050
+SI	Conditional	0.0003	92.6	0.0188	0.9061	0.0053
	Marginal	0.0003	91.4	0.0188	0.8913	0.0054
+MI; 5	Conditional	-0.0000	91.4	0.0136	0.8935	0.0052
	Marginal	0.0003	93.8	0.0196	0.9537	0.0052
+MI; 10	Conditional	0.0003	94.8	0.0197	0.9643	0.0052
	Marginal	0.0003	93.8	0.0196	0.9537	0.0052
+MI-B; 5	Conditional	0.0007	94.2	0.0203	0.9785	0.0054
	Marginal	0.0007	94.6	0.0202	0.9725	0.0054
+MI-B; 10	Conditional	0.0007	94.0	0.0202	0.9814	0.0053
	Marginal	0.0007	95.2	0.0201	0.9740	0.0053
$Q_1 = 2$						
HMM		0.0006	94.8	0.0193	0.9585	0.0051
+SI	Conditional	0.0003	91.6	0.0188	0.8988	0.0053
	Marginal	0.0002	93.8	0.0188	0.9261	0.0052
+MI; 5	Conditional	0.0004	94.0	0.0230	0.9822	0.0051
	Marginal	0.0002	94.6	0.0196	0.9867	0.0051
+MI; 10	Conditional	0.0002	95.0	0.0199	0.9886	0.0051
	Marginal	0.0002	94.6	0.0196	0.9867	0.0051
+MI-B; 5	Conditional	0.0006	95.6	0.0206	1.0273	0.0052
	Marginal	0.0006	96.0	0.0205	1.0178	0.0052
+MI-B; 10	Conditional	0.0006	96.6	0.0206	1.0333	0.0051
	Marginal	0.0006	95.8	0.0203	1.0191	0.0051
$Q_1 = 3$						
HMM		0.0002	92.2	0.0194	0.8818	0.0056
+SI	Conditional	0.0000	88.0	0.0188	0.8071	0.0059
	Marginal	-0.0001	89.8	0.0188	0.8296	0.0058
+MI; 5	Conditional	0.0005	94.4	0.0493	0.9838	0.0057
	Marginal	0.0000	92.4	0.0197	0.8878	0.0057
+MI; 10	Conditional	-0.0000	93.8	0.0200	0.8968	0.0057
	Marginal	0.0000	92.4	0.0197	0.8878	0.0057
+MI-B; 5	Conditional	0.0004	92.6	0.0209	0.9317	0.0057
	Marginal	0.0004	93.0	0.0207	0.9194	0.0058
+MI-B; 10	Conditional	0.0004	93.6	0.0208	0.9254	0.0057
	Marginal	0.0004	93.2	0.0205	0.9177	0.0057

Appendix D

Table D1: All simulation results obtained when no correction method was applied. The table shows results for bias of the logit coefficient, coverage of the 95% confidence interval, ratio of the average standard error of the estimate over the standard deviation of the $1,000$ replication estimates and root mean squared error. Results are shown using sample sizes of 200, 500 and , 1000. Logit coefficients for covariate Q_1 are 0.50, 0.62, 0.73 and 0.88. Proportions of $P(Q_2) = 2$ are 0.01, 0.05, 0.10 and 0.20.

			0.70				0.80				0.90				0.95				0.99			
N	Q_1	Q_2	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse
200	0.50	0.01	-0.0049	0.9970	1.7283	0.2441	0.0007	0.9930	1.3436	0.2535	-0.0029	0.9680	1.0946	0.2775	0.0063	0.9650	1.0435	0.2794	0.0048	0.9560	1.0063	0.2850
		0.05	0.0172	0.9980	1.7475	0.2400	0.0034	0.9890	1.3262	0.2566	0.0095	0.9690	1.1099	0.2738	0.0020	0.9480	1.0148	0.2870	-0.0002	0.9520	1.0128	0.2831
		0.10	0.0062	0.9970	1.6622	0.2552	-0.0067	0.9950	1.3882	0.2450	0.0007	0.9650	1.0913	0.2780	0.0008	0.9660	1.0332	0.2819	0.0015	0.9520	0.9943	0.2884
		0.20	-0.0032	0.9970	1.6936	0.2457	0.0093	0.9900	1.3944	0.2449	0.0109	0.9730	1.1160	0.2725	0.0090	0.9550	1.0008	0.2912	-0.0096	0.9610	1.0320	0.2780
	0.62	0.01	0.2900	0.9760	1.7343	0.3783	0.1623	0.9780	1.4050	0.2949	0.0632	0.9660	1.0827	0.2909	-0.0024	0.9630	1.0381	0.2854	0.0007	0.9480	1.0097	0.2886
		0.05	0.2917	0.9820	1.6738	0.3871	0.1722	0.9780	1.4153	0.2988	0.0475	0.9610	1.0646	0.2942	0.0119	0.9700	1.0604	0.2793	0.0029	0.9560	1.0300	0.2829
		0.10	0.2817	0.9700	1.6530	0.3810	0.1623	0.9740	1.3702	0.3017	0.0571	0.9550	1.0898	0.2875	-0.0013	0.9670	1.0276	0.2881	0.0028	0.9530	1.0223	0.2851
		0.20	0.2923	0.9740	1.6829	0.3876	0.1655	0.9780	1.3353	0.3065	0.0475	0.9700	1.0984	0.2840	0.0103	0.9620	1.0376	0.2857	-0.0018	0.9620	1.0545	0.2763
	0.73	0.01	0.5591	0.8278	1.7595	0.6133	0.3195	0.9160	1.3569	0.4155	0.1159	0.9510	1.0937	0.3173	0.0327	0.9470	1.0099	0.3100	0.0056	0.9570	1.0304	0.2968
		0.05	0.5705	0.8220	1.6922	0.6257	0.3483	0.8930	1.3515	0.4387	0.0840	0.9590	1.0763	0.3122	0.0106	0.9490	1.0031	0.3105	-0.0069	0.9400	0.9724	0.3151
		0.10	0.5980	0.8010	1.6946	0.6516	0.3282	0.9080	1.3216	0.4267	0.1020	0.9590	1.1048	0.3098	0.0196	0.9610	1.0415	0.2997	0.0048	0.9570	1.0150	0.3016
		0.20	0.5779	0.8000	1.6575	0.6369	0.3317	0.9200	1.3539	0.4249	0.0906	0.9690	1.1823	0.2879	0.0432	0.9570	1.0287	0.3054	-0.0027	0.9510	1.0087	0.3042
	0.88	0.01	1.1947	0.2422	1.7489	1.2285	0.7421	0.5780	1.4087	0.7993	0.2467	0.9110	1.1316	0.4228	0.0516	0.9650	1.0541	0.3672	0.0098	0.9650	1.0469	0.3593
		0.05	1.2018	0.2377	1.7746	1.2340	0.7292	0.6000	1.3616	0.7911	0.2209	0.9250	1.1173	0.4155	0.0403	0.9560	1.0010	0.3871	-0.0184	0.9600	1.0127	0.3748
		0.10	1.2022	0.2324	1.7425	1.2351	0.7470	0.5610	1.3522	0.8084	0.2533	0.9230	1.1363	0.4254	0.0523	0.9500	1.0255	0.3779	-0.0055	0.9620	1.0393	0.3634
		0.20	1.2002	0.2404	1.7385	1.2335	0.7252	0.5720	1.3279	0.7901	0.2448	0.9110	1.1243	0.4247	0.0484	0.9660	1.0532	0.3669	-0.0342	0.9640	0.9927	0.3847
500	0.50	0.01	0.0075	1.0000	1.8485	0.1409	0.0026	0.9950	1.3551	0.1572	-0.0005	0.9570	1.0475	0.1817	-0.0156	0.9650	1.0474	0.1750	0.0042	0.9460	0.9998	0.1802
		0.05	0.0024	0.9980	1.8366	0.1398	-0.0075	0.9950	1.3875	0.1542	-0.0034	0.9640	1.0787	0.1768	0.0034	0.9450	1.0009	0.1825	-0.0112	0.9540	1.0082	0.1790
		0.10	-0.0034	0.9970	1.7360	0.1482	0.0137	0.9870	1.3679	0.1555	-0.0028	0.9680	1.1019	0.1730	-0.0020	0.9530	1.0144	0.1799	-0.0012	0.9510	1.0291	0.1750
		0.20	0.0047	0.9980	1.7682	0.1458	0.0013	0.9970	1.4219	0.1499	-0.0059	0.9700	1.1092	0.1716	0.0021	0.9660	1.0365	0.1761	-0.0035	0.9610	0.9953	0.1810
	0.62	0.01	0.2954	0.8880	1.7821	0.3293	0.1591	0.9490	1.3595	0.2253	0.0560	0.9620	1.1147	0.1823	0.0141	0.9420	0.9970	0.1865	0.0076	0.9610	1.0439	0.1754
		0.05	0.2994	0.9060	1.7942	0.3326	0.1745	0.9420	1.3494	0.2371	0.0416	0.9730	1.1408	0.1743	0.0032	0.9480	0.9985	0.1859	0.0039	0.9510	1.0035	0.1824
		0.10	0.2918	0.9010	1.7108	0.3291	0.1722	0.9360	1.3691	0.2339	0.0527	0.9600	1.0911	0.1848	0.0114	0.9610	1.0326	0.1800	0.0058	0.9580	0.9971	0.1836
		0.20	0.3008	0.8980	1.7803	0.3344	0.1698	0.9300	1.3641	0.2326	0.0435	0.9710	1.1488	0.1737	0.0084	0.9580	1.0154	0.1829	-0.0002	0.9620	1.0364	0.1766
	0.73	0.01	0.6007	0.3410	1.7687	0.6197	0.3451	0.7040	1.3651	0.3816	0.1057	0.9400	1.1398	0.2069	0.0345	0.9530	1.0454	0.1894	-0.0005	0.9610	1.0002	0.1921
		0.05	0.5962	0.3110	1.7283	0.6159	0.3463	0.6930	1.3429	0.3842	0.1029	0.9320	1.1134	0.2091	0.0290	0.9690	1.0797	0.1824	0.0008	0.9580	0.9959	0.1930
		0.10	0.5986	0.3280	1.7121	0.6188	0.3522	0.6850	1.4410	0.3849	0.1024	0.9340	1.1151	0.2085	0.0288	0.9480	1.0128	0.1942	0.0056	0.9560	1.0233	0.1877
		0.20	0.6048	0.2990	1.8260	0.6224	0.3317	0.7350	1.3867	0.3691	0.1059	0.9430	1.1125	0.2106	0.0287	0.9530	1.0385	0.1894	-0.0062	0.9530	0.9981	0.1928
	0.88	0.01	1.2516	0.0170	1.7013	1.2639	0.7561	0.0990	1.4520	0.7764	0.2564	0.8300	1.1290	0.3346	0.0614	0.9500	1.0294	0.2390	-0.0072	0.9520	0.9930	0.2377
		0.05	1.2527	0.0150	1.7392	1.2645	0.7497	0.1120	1.4379	0.7707	0.2498	0.8240	1.1035	0.3330	0.0445	0.9420	0.9873	0.2460	0.0083	0.9590	1.0140	0.2323
		0.10	1.2461	0.0140	1.7734	1.2576	0.7549	0.1170	1.4604	0.7749	0.2477	0.8280	1.1366	0.3268	0.0685	0.9580	1.0563	0.2353	0.0050	0.9540	1.0119	0.2322
		0.20	1.2635	0.0170	1.7066	1.2756	0.7625	0.1070	1.4326	0.7836	0.2387	0.8360	1.0845	0.3279	0.0460	0.9420	1.0035	0.2426	-0.0073	0.9420	0.9904	0.2387
	0.50	0.01	-0.0004	0.9980	1.7567	0.0994	-0.0018	0.9950	1.3605	0.1103	0.0026	0.9720	1.1275	0.1191	-0.0044	0.9650	1.0629	0.1212	-0.0025	0.9570	0.9856	0.1287
		0.05	-0.0018	1.0000	1.8566	0.0947	0.0017	0.9930	1.4154	0.1054	-0.0042	0.9770	1.1146	0.1202	0.0057	0.9610	1.0442	0.1235	-0.0012	0.9420	0.9774	0.1298
		0.10	-0.0032	1.0000	1.8534	0.0953	0.0016	0.9910	1.4325	0.1051	-0.0034	0.9690	1.1031	0.1216	0.0013	0.9570	1.0367	0.1243	-0.0023	0.9480	0.9561	0.1327
		0.20	-0.0022	0.9980	1.7579	0.0998	-0.0015	0.9920	1.3711	0.1093	0.0014	0.9730	1.0903	0.1229	0.0039	0.9550	1.0027	0.1285	-0.0002	0.9430	1.0074	0.1259
	0.62	0.01	0.3049	0.6410	1.7658	0.3211	0.1683	0.8790	1.4083	0.2002	0.0506	0.9560	1.0977	0.1339	0.0132	0.9590	1.0641	0.1236	0.0046	0.9530	1.0256	0.1257
		0.05	0.3011	0.6460	1.8005	0.3169	0.1730	0.8540	1.3935	0.2049	0.0528	0.9460	1.1366	0.1308	0.0095	0.9500	1.0147	0.1293	0.0023	0.9590	1.0146	0.1273
		0.10	0.3077	0.6260	1.8357	0.3227	0.1720	0.8820	1.4213	0.2029	0.0514	0.9490	1.0897	0.1351	0.0100	0.9520	1.0106	0.1298	-0.0053	0.9400	0.9609	0.1345
		0.20	0.3066	0.6480	1.8116	0.3221	0.1731	0.8680	1.3822	0.2051	0.0526	0.9530	1.1361	0.1307	0.0117	0.9500	1.0338	0.1272	-0.0004	0.9570	1.0103	0.1279
1000	0.73	0.01	0.6198	0.0220	1.8220	0.6279	0.3509	0.3780	1.3041	0.3711	0.1005	0.9210	1.1051	0.1633	0.0292	0.9520	1.0192	0.1377	0.0027	0.9420	0.9757	0.1386
		0.05	0.6127	0.0330	1.8018	0.6212	0.3519	0.3620	1.4517	0.3685	0.1055	0.9100	1.0899	0.1683	0.0281	0.9560	1.0376	0.1353	0.0093	0.9480	1.0020	0.1352
		0.10	0.6135	0.0250	1.8031	0.6219	0.3373	0.4130	1.3994	0.3556	0.0999	0.9060	1.0907	0.1644	0.0268	0.9560	1.0575	0.1325	0.0032	0.9610	1.0269	0.1317
		0.20	0.6130	0.0240	1.8021	0.6214	0.3470	0.3890	1.3817	0.3652	0.1059	0.9210	1.1429	0.1636	0.0302	0.9500	1.0281	0.1368	0.0044	0.9450	1.0076	0.1342
	0.88	0.01	1.2822	0.0000	1.7734	1.2873	0.7610	0.0010	1.3578	0.7723	0.2430	0.7080	1.0693	0.2909	0.0697	0.9380	1.0258	0.1772	0.0023	0.9440	0.9778	0.1691
		0.05	1.2897	0.0000	1.7444	1.2949	0.7661	0.0010	1.3900	0.7770	0.2404	0.7270	1.1445	0.2831	0.0645	0.9390	1.0476	0.1722	-0.0038	0.9550	0.9956	0.1663
		0.10	1.2914	0.0000	1.7953	1.2963	0.7651	0.0040	1.4133	0.7756	0.2342	0.7350	1.1495	0.2775	0.0652	0.9480	1.0473	0.1725	0.0029	0.9600	1.0357	0.1596
		0.20	1.2900	0.0010	1.7073	1.2955	0.7661	0.0030	1.4238	0.7764	0.2473	0.7080	1.1572	0.2880	0.0707	0.9390	1.0797	0.1697	0.0041	0.9580	1.0319	0.1605

Table D2: All simulation results obtained when the maximum likelihood three-step correction procedure was applied. The table shows results for bias of the logit coefficient, coverage of the 95% confidence interval, ratio of the average standard error of the estimate over the standard deviation of the 1,000 replication estimates and root mean squared error. Results are shown using sample sizes of 200, 500 and , 1000. Logit coefficients for covariate Q_1 are 0.50, 0.62, 0.73 and 0.88. Proportions of $P(Q_2) = 2$ are 0.01, 0.05, 0.10 and 0.20.

			0.70				0.80				0.90				0.95				0.99			
N	Q_1	Q_2	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse
		0.01	-0.0124	0.9820	1.3137	0.4360	0.0059	0.9610	1.1033	0.3570	-0.0028	0.9640	1.0418	0.3017	0.0018	0.9610	0.9896	0.2883	0.0037	0.9580	1.0206	0.2898
		0.05	0.0404	0.9810	1.3917	0.4644	0.0052	0.9620	1.1378	0.3866	0.0117	0.9700	1.0758	0.3150	0.0050	0.9490	1.0014	0.3083	0.0003	0.9520	1.0067	0.2893
	0.0	0.10	-0.0054	0.9830	1.3197	0.4704	-0.0040	0.9720	1.1900	0.3607	0.0025	0.9510	1.0393	0.3221	0.0018	0.9530	1.0111	0.3027	0.0019	0.9510	0.9909	0.2935
		0.20	0.0019	0.9780	1.2831	0.4522	0.0151	0.9680	1.1638	0.3559	0.0116	0.9700	1.0674	0.3094	0.0102	0.9510	0.9857	0.3082	-0.0091	0.9620	1.0301	0.2815
		0.01	0.1795	0.8450	1.1580	0.5415	0.0894	0.8640	1.0026	0.4243	0.0988	0.8340	0.8007	0.4117	0.0587	0.8360	0.7585	0.3954	0.0100	0.9400	1.0101	0.3130
		0.05	0.1025	0.9790	1.3439	0.4971	0.0193	0.9690	1.1849	0.3763	-0.0129	0.9500	1.0202	0.3380	-0.0198	0.9630	1.0429	0.3010	-0.0062	0.9550	1.0242	0.2896
	0.5	0.10	0.0801	0.9760	1.3095	0.4904	-0.0027	0.9650	1.1868	0.3770	-0.0075	0.9530	1.0478	0.3252	-0.0323	0.9580	1.0158	0.3092	-0.0050	0.9510	1.0181	0.2917
		0.20	0.0804	0.9610	1.2650	0.4782	0.0035	0.9680	1.1366	0.3779	-0.0154	0.9620	1.0691	0.3138	-0.0165	0.9580	1.0238	0.3023	-0.0074	0.9590	1.0538	0.2805
200		0.01	0.3222	0.8400	1.0170	0.7261	0.1811	0.8350	0.7005	0.6567	0.1953	0.8240	0.5500	0.6418	0.1427	0.8300	0.5326	0.6093	0.0283	0.9450	0.9123	0.3894
		0.05	0.1827	0.9580	1.3479	0.5356	0.0373	0.9700	1.1736	0.4155	-0.0483	0.9590	1.0386	0.3562	-0.0615	0.9460	0.9578	0.3546	-0.0265	0.9420	0.9754	0.3241
	1.0	0.10	0.2200	0.9530	1.3500	0.5472	-0.0005	0.9710	1.1213	0.4220	-0.0258	0.9520	1.0385	0.3491	-0.0439	0.9580	1.0201	0.3291	-0.0105	0.9580	1.0151	0.3085
		0.20	0.1739	0.9600	1.2946	0.5201	0.0084	0.9720	1.1771	0.3892	-0.0369	0.9710	1.1141	0.3226	-0.0134	0.9630	1.0111	0.3245	-0.0173	0.9550	1.0004	0.3134
		0.01	0.7106	0.8049	0.8475	1.2076	0.3811	0.8504	0.5629	1.1368	0.3159	0.8257	0.4066	1.1509	0.2043	0.8521	0.4016	1.0875	0.0310	0.9659	0.8793	0.5611
		0.05	0.3959	0.9088	1.4372	0.7427	0.0084	0.9646	1.2142	0.5421	-0.1181	0.9719	1.0481	0.4950	-0.1556	0.9620	0.9620	0.4924	-0.0705	0.9660	1.0160	0.4079
	2.0	0.10	0.3783	0.9116	1.4199	0.7151	0.0355	0.9549	1.1623	0.5488	-0.0748	0.9710	1.0125	0.4810	-0.1394	0.9570	0.9450	0.4942	-0.0567	0.9720	1.0310	0.3989
		0.20	0.4120	0.9168	1.4320	0.6867	0.0138	0.9730	1.1882	0.4989	-0.0633	0.9588	1.0308	0.4572	-0.1056	0.9729	1.0113	0.4383	-0.0784	0.9670	0.9980	0.4124
		0.01	0.0143	0.9850	1.4363	0.3042	0.0053	0.9680	1.1461	0.2456	0.0002	0.9460	1.0169	0.2077	-0.0162	0.9680	1.0393	0.1859	0.0036	0.9490	1.0053	0.1832
		0.05	0.0057	0.9820	1.3805	0.3058	-0.0091	0.9700	1.1867	0.2352	-0.0029	0.9570	1.0523	0.2008	0.0032	0.9440	0.9950	0.1920	-0.0112	0.9540	1.0106	0.1818
	0.0	0.10	-0.0041	0.9740	1.3159	0.3124	0.0205	0.9690	1.1522	0.2359	-0.0020	0.9600	1.0722	0.1957	-0.0025	0.9530	1.0109	0.1876	-0.0010	0.9510	1.0284	0.1778
		0.20	0.0103	0.9700	1.3433	0.2906	0.0045	0.9780	1.2065	0.2217	-0.0060	0.9650	1.0766	0.1909	0.0008	0.9550	1.0307	0.1830	-0.0038	0.9580	0.9929	0.1842
		0.01	0.0609	0.9570	1.2991	0.3366	-0.0119	0.9600	1.1626	0.2489	0.0059	0.9580	1.0213	0.2102	-0.0089	0.9410	0.9660	0.2017	0.0001	0.9660	1.0138	0.1848
		0.05	0.0624	0.9700	1.3612	0.3266	0.0115	0.9650	1.1444	0.2504	-0.0154	0.9610	1.1137	0.1931	-0.0205	0.9430	0.9902	0.1971	-0.0056	0.9520	1.0026	0.1864
	0.5	0.10	0.0432	0.9660	1.2986	0.3304	0.0051	0.9650	1.1837	0.2387	-0.0030	0.9580	1.0595	0.2011	-0.0111	0.9550	1.0187	0.1896	-0.0024	0.9540	0.9943	0.1877
		0.20	0.0516	0.9740	1.3534	0.3044	0.0018	0.9650	1.1396	0.2382	-0.0130	0.9580	1.0899	0.1926	-0.0111	0.9550	1.0013	0.1922	-0.0070	0.9620	1.0367	0.1796
500		0.01	0.1323	0.9630	1.3100	0.3844	0.0145	0.9620	1.0609	0.2823	-0.0090	0.9640	0.9754	0.2335	-0.0041	0.9370	0.8763	0.2341	-0.0127	0.9540	0.8538	0.2368
		0.05	0.1102	0.9670	1.3694	0.3583	0.0070	0.9620	1.1755	0.2588	-0.0157	0.9660	1.0839	0.2109	-0.0193	0.9630	1.0609	0.1941	-0.0173	0.9550	1.0025	0.1978
	1.0	0.10	0.1018	0.9610	1.3112	0.3639	0.0160	0.9730	1.2315	0.2439	-0.0146	0.9540	1.0727	0.2107	-0.0167	0.9510	0.9956	0.2056	-0.0138	0.9550	1.0238	0.1933
		0.20	0.1070	0.9560	1.3462	0.3388	-0.0161	0.9670	1.1665	0.2461	-0.0103	0.9570	1.0585	0.2103	-0.0149	0.9570	1.0162	0.2002	-0.0229	0.9500	0.9952	0.1993
		0.01	0.2582	0.9188	1.1959	0.6039	-0.0002	0.9700	1.0471	0.3907	-0.0175	0.9630	0.8218	0.3585	-0.0684	0.9480	0.8339	0.3247	-0.0645	0.9460	0.8116	0.3307
		0.05	0.2297	0.9310	1.3481	0.5199	-0.0339	0.9620	1.1286	0.3670	-0.0425	0.9630	1.0360	0.2895	-0.0865	0.9270	0.9279	0.2995	-0.0484	0.9510	0.9642	0.2683
	2.0	0.10	0.2202	0.9440	1.4179	0.4925	-0.0163	0.9710	1.1827	0.3344	-0.0436	0.9710	1.0793	0.2751	-0.0508	0.9590	1.0005	0.2660	-0.0546	0.9550	0.9662	0.2710
		0.20	0.2596	0.9150	1.3026	0.5061	0.0159	0.9630	1.1932	0.3205	-0.0436	0.9590	1.0257	0.2817	-0.0681	0.9520	0.9726	0.2776	-0.0624	0.9410	0.9592	0.2719
		0.01	0.0022	0.9890	1.4474	0.1672	-0.0030	0.9680	1.1558	0.1730	0.0022	0.9620	1.0844	0.1367	-0.0056	0.9610	1.0665	0.1261	-0.0028	0.9560	0.9861	0.1308
		0.05	-0.0028	0.9750	1.3669	0.2271	0.0017	0.9730	1.1982	0.1629	-0.0056	0.9680	1.0793	0.1376	0.0051	0.9580	1.0339	0.1297	-0.0013	0.9420	0.9759	0.1316
	0.0	0.10	-0.0103	0.9820	1.3954	0.2200	0.0040	0.9750	1.2109	0.1623	-0.0038	0.9640	1.0757	0.1374	0.0005	0.9550	1.0339	0.1290	-0.0024	0.9470	0.9588	0.1339
		0.20	0.0026	0.9680	1.3071	0.2220	0.0010	0.9680	1.1573	0.1625	0.0002	0.9600	1.0546	0.1371	0.0041	0.9490	0.9934	0.1333	0.0001	0.9440	1.0050	0.1280
		0.01	0.0343	0.9680	1.3147	0.2479	-0.0026	0.9670	1.2081	0.1693	-0.0052	0.9640	1.0721	0.1416	-0.0074	0.9560	1.0568	0.1295	-0.0033	0.9560	1.0199	0.1288
		0.05	0.0279	0.9710	1.3324	0.2396	0.0075	0.9700	1.1811	0.1723	0.0304	0.9620	1.1246	0.1298	-0.0104	0.9480	1.0116	0.1353	-0.0077	0.9590	1.0128	0.1306
	0.5	0.10	0.0336	0.9650	1.3693	0.2309	0.0054	0.9710	1.2052	0.1667	0.0275	0.9570	1.0779	0.1338	-0.0096	0.9510	1.0085	0.1350	-0.0154	0.9350	0.9591	0.1382
		0.20	0.0285	0.9720	1.3482	0.2246	0.0027	0.9680	1.1606	0.1666	0.0285	0.9670	1.1187	0.1281	-0.0061	0.9490	1.0160	0.1326	-0.0092	0.9500	0.9998	0.1316
1000		0.01	0.0762	0.9650	1.3840	0.2632	0.0088	0.9580	1.1202	0.1938	-0.0144	0.9620	1.0659	0.1500	-0.0126	0.9530	1.0051	0.1432	-0.0157	0.9440	0.9734	0.1435
		0.05	0.0629	0.9710	1.4037	0.2580	0.0097	0.9820	1.2510	0.1725	-0.0078	0.9590	1.0502	0.1525	-0.0132	0.9590	1.0322	0.1398	-0.0038	0.9480	0.9950	0.1385
	1.0	0.10	0.0564	0.9590	1.3270	0.2577	-0.0153	0.9680	1.1670	0.1825	-0.0127	0.9580	1.0528	0.1508	-0.0112	0.9620	1.0484	0.1369	-0.0107	0.9580	1.0210	0.1357
		0.20	0.0490	0.9630	1.3576	0.2409	-0.0021	0.9590	1.1493	0.1765	-0.0053	0.9630	1.0756	0.1451	-0.0050	0.9520	1.0134	0.1402	-0.0095	0.9490	0.9948	0.1390
		0.01	0.1389	0.9480	1.4218	0.3662	-0.0181	0.9650	1.1043	0.2559	-0.0304	0.9580	0.9945	0.2083	-0.0364	0.9510	0.9609	0.1917	-0.0445	0.9480	0.9543	0.1891
		0.05	0.1341	0.9420	1.3666	0.3766	-0.0064	0.9560	1.1348	0.2475	-0.0302	0.9620	1.0679	0.1938	-0.0343	0.9510	0.9992	0.1843	-0.0466	0.9540	0.9661	0.1889
	2.0	0.10	0.1366	0.9320	1.3274	0.3724	-0.0037	0.9640	1.1757	0.2401	-0.0376	0.9620	1.0926	0.1899	-0.0319	0.9610	0.9977	0.1840	-0.0413	0.9470	1.0059	0.1807
		0.20	0.1434	0.9380	1.3617	0.3445	0.0031	0.9660	1.1499	0.2304	-0.0187	0.9660	1.0766	0.1856	-0.0207	0.9630	1.0467	0.1722	-0.0350	0.9550	1.0038	0.1778

Table D3: All simulation results obtained when the BCH three-step correction procedure was applied. The table shows results for bias of the logit coefficient, coverage of the 95% confidence interval, ratio of the average standard error of the estimate over the standard deviation of the $1,000$ replication estimates and root mean squared error. Results are shown using sample sizes of 200, 500 and , 1000. Logit coefficients for covariate Q_1 are 0.50, 0.62, 0.73 and 0.88. Proportions of $P(Q_2) = 2$ are 0.01, 0.05, 0.10 and 0.20.

			0.70				0.80				0.90				0.95				0.99			
N	Q	Z	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse	bias	cov	sesd	rmse
		0.01	0.0076	0.9956	1.5168	0.4692	-0.0030	0.9729	1.1456	0.4247	-0.0006	0.9735	1.0892	0.3112	0.0047	0.9632	1.0245	0.2939	0.0049	0.9580	0.9973	0.2878
		0.05	0.0373	0.9951	1.4356	0.4946	0.0047	0.9717	1.1321	0.4280	0.0107	0.9640	1.0628	0.3131	0.0046	0.9524	1.0121	0.2966	0.0046	0.9539	1.0085	0.2845
	0.0	0.10	-0.0347	0.9924	1.4598	0.4254	0.0025	0.9731	1.1713	0.3913	0.0005	0.9609	1.0587	0.3147	0.0009	0.9575	1.0347	0.2898	0.0019	0.9514	1.0062	0.2848
		0.20	-0.0190	0.9870	1.3555	0.3856	0.0081	0.9732	1.1931	0.3641	0.0100	0.9678	1.0623	0.3077	0.0108	0.9536	0.9897	0.3010	0.0032	0.9647	1.0421	0.2746
		0.01	0.1382	0.9502	1.4826	0.5027	-0.0397	0.9810	1.2583	0.4094	0.0121	0.9606	1.0631	0.3223	-0.0206	0.9575	1.0208	0.3012	0.0032	0.9471	1.0069	0.2895
	0.5	0.05	0.1072	0.9836	1.4926	0.5051	-0.0125	0.9791	1.2189	0.4055	-0.0099	0.9628	1.0395	0.3289	-0.0024	0.9689	1.0466	0.2918	0.0102	0.9527	1.0179	0.2864
		0.10	0.1072	0.9745	1.4551	0.4593	-0.0275	0.9699	1.1715	0.4165	0.0081	0.9608	1.0567	0.3182	-0.0183	0.9666	1.0238	0.2982	-0.0067	0.9535	1.0181	0.2864
		0.20	0.1377	0.9603	1.3741	0.4243	-0.0055	0.9669	1.1407	0.3893	-0.0092	0.9648	1.0706	0.3115	-0.0043	0.9619	1.0324	0.2934	0.0025	0.9400	0.9827	0.2961
200		0.01	0.2829	0.8989	1.4514	0.5767	-0.0597	0.9787	1.1550	0.4903	0.0046	0.9603	1.0493	0.3471	-0.0007	0.9503	1.0212	0.3164	0.0082	0.9571	1.0344	0.2956
		0.05	0.1874	0.9326	1.4495	0.5648	-0.0322	0.9694	1.1663	0.4768	-0.0318	0.9610	1.0520	0.3474	-0.0211	0.9525	1.0017	0.3227	-0.0122	0.9356	0.9617	0.3190
	1.0	0.10	0.2653	0.9197	1.4250	0.5410	-0.0433	0.9612	1.1162	0.4767	-0.0037	0.9597	1.0691	0.3359	-0.0126	0.9582	1.0340	0.3113	-0.0008	0.9578	1.0116	0.3029
		0.20	0.2496	0.9279	1.4216	0.4787	0.0005	0.9783	1.1912	0.4013	-0.0169	0.9759	1.1254	0.3124	0.0090	0.9652	1.0255	0.3122	0.0095	0.9461	0.9820	0.3119
		0.01	0.8146	0.7244	1.3904	0.9743	0.0740	0.9551	1.2614	0.5485	-0.0305	0.9647	1.0838	0.4474	-0.0380	0.9610	1.0342	0.4019	0.0019	0.9639	1.0566	0.3567
		0.05	0.6159	0.8169	1.4368	0.8763	-0.0381	0.9645	1.1828	0.6150	-0.0606	0.9719	1.0534	0.4684	-0.0464	0.9626	0.9924	0.4219	-0.0187	0.9597	1.0082	0.3768
	2.0	0.10	0.4809	0.8754	1.3606	0.7923	-0.0335	0.9563	1.1696	0.6044	-0.0181	0.9758	1.0766	0.4404	-0.0275	0.9582	1.0342	0.3975	-0.0076	0.9603	1.0457	0.3614
		0.20	0.4407	0.8977	1.3971	0.6840	0.0340	0.9703	1.1833	0.5154	-0.0209	0.9649	1.0729	0.4353	-0.0331	0.9724	1.0297	0.3969	-0.0177	0.9664	1.0209	0.3703
		0.01	-0.0020	0.9921	1.4683	0.4196	0.0011	0.9699	1.1589	0.2573	-0.0025	0.9458	1.0167	0.2062	-0.0162	0.9650	1.0407	0.1820	0.0034	0.9479	0.9977	0.1808
		0.05	-0.0247	0.9889	1.5482	0.3270	-0.0102	0.9640	1.1537	0.2507	-0.0045	0.9560	1.0464	0.2011	0.0043	0.9420	0.9996	0.1885	-0.0098	0.9479	0.9950	0.1814
	0.0	0.10	-0.0141	0.9814	1.4250	0.3157	0.0151	0.9540	1.1355	0.2483	-0.0020	0.9590	1.0775	0.1928	-0.0017	0.9570	1.0130	0.1855	-0.0018	0.9503	1.0171	0.1772
		0.20	0.0101	0.9888	1.4966	0.2675	0.0036	0.9730	1.1670	0.2325	-0.0052	0.9630	1.0760	0.1910	0.0014	0.9610	1.0257	0.1819	-0.0050	0.9606	0.9952	0.1811
		0.01	-0.0359	0.9121	1.4246	0.4576	-0.0415	0.9648	1.1400	0.2795	0.0024	0.9697	1.0952	0.1956	0.0003	0.9418	1.0019	0.1911	0.0078	0.9664	1.0479	0.1749
		0.05	0.0537	0.9435	1.4133	0.3908	-0.0131	0.9619	1.1311	0.2706	-0.0137	0.9730	1.1034	0.1930	-0.0119	0.9460	0.9918	0.1933	0.0029	0.9494	1.0019	0.1829
	0.5	0.10	0.0787	0.9253	1.3717	0.3481	-0.0158	0.9710	1.1650	0.2554	0.0003	0.9540	1.0499	0.2023	-0.0053	0.9600	1.0250	0.1867	0.0057	0.9574	0.9943	0.1843
		0.20	0.0938	0.9426	1.4273	0.2958	-0.0149	0.9700	1.1148	0.2490	-0.0118	0.9670	1.0994	0.1905	-0.0069	0.9580	1.0138	0.1875	0.0002	0.9654	1.0270	0.1783
500		0.01	0.0113	0.9303	1.4137	0.5542	-0.0556	0.9684	1.1216	0.3100	-0.0055	0.9719	1.1092	0.2052	0.0036	0.9547	1.0455	0.1930	-0.0001	0.9633	1.0072	0.1910
		0.05	0.0695	0.9256	1.4048	0.4637	-0.0468	0.9649	1.0919	0.3059	-0.0069	0.9610	1.0770	0.2096	-0.0019	0.9690	1.0714	0.1880	-0.0008	0.9600	0.9927	0.1938
	1.0	0.10	0.1237	0.8855	1.3289	0.4153	-0.0169	0.9760	1.1780	0.2676	-0.0067	0.9570	1.0814	0.2077	-0.0033	0.9540	1.0115	0.1987	0.0069	0.9530	1.0161	0.1892
		0.20	0.1604	0.9123	1.4270	0.3581	-0.0395	0.9710	1.1426	0.2647	-0.0042	0.9650	1.0717	0.2066	-0.0034	0.9550	1.0360	0.1932	-0.0140	0.9526	0.9880	0.1953
		0.01	0.4147	0.8061	1.1017	0.8573	-0.1014	0.9747	1.1125	0.4763	-0.0100	0.9668	1.0620	0.2783	-0.0172	0.9549	1.0291	0.2473	-0.0100	0.9511	0.9912	0.2389
		0.05	0.1464	0.8714	1.1630	0.7303	-0.1594	0.9560	0.9934	0.5423	-0.0182	0.9560	1.0419	0.2842	-0.0357	0.9540	0.9757	0.2637	0.0046	0.9633	1.0184	0.2317
	2.0	0.10	0.2508	0.8798	1.2883	0.6013	-0.1318	0.9487	1.0008	0.4843	-0.0170	0.9750	1.0925	0.2691	-0.0094	0.9650	1.0393	0.2434	-0.0011	0.9544	0.9988	0.2360
		0.20	0.3738	0.8368	1.2830	0.5731	-0.0330	0.9540	1.1122	0.3775	-0.0330	0.9540	1.0237	0.2832	-0.0322	0.9580	1.0013	0.2542	-0.0131	0.9431	0.9983	0.2375
		0.01	-0.0004	0.9942	1.3762	0.3841	-0.0067	0.9630	1.1709	0.1748	0.0020	0.9710	1.1061	0.1341	-0.0054	0.9670	1.0555	0.1262	-0.0027	0.9540	0.9852	0.1290
		0.05	-0.0063	0.9873	1.4482	0.2672	-0.0005	0.9700	1.2060	0.1656	-0.0054	0.9640	1.0789	0.1366	0.0067	0.9560	1.0363	0.1281	-0.0009	0.9409	0.9768	0.1301
	0.0	0.10	-0.0137	0.9828	1.3935	0.2578	0.0069	0.9700	1.2170	0.1637	-0.0039	0.9630	1.0669	0.1371	0.0015	0.9550	1.0307	0.1284	-0.0028	0.9474	0.9572	0.1327
		0.20	0.0095	0.9719	1.3477	0.2348	-0.0003	0.9620	1.1582	0.1642	0.0011	0.9640	1.0547	0.1367	0.0045	0.9560	0.9936	0.1326	-0.0000	0.9425	1.0108	0.1256
		0.01	-0.1471	0.9773	1.5598	0.4322	-0.0167	0.9610	1.1787	0.1787	-0.0017	0.9610	1.0719	0.1407	-0.0015	0.9600	1.0602	0.1273	0.0040	0.9550	1.0263	0.1258
		0.05	-0.0349	0.9254	1.4313	0.3258	-0.0038	0.9710	1.1791	0.1755	0.0016	0.9650	1.0956	0.1365	-0.0047	0.9440	1.0134	0.1332	-0.0009	0.9590	1.0134	0.1284
	0.5	0.10	0.0307	0.9302	1.4639	0.2734	-0.0048	0.9700	1.2062	0.1697	-0.0003	0.9550	1.0574	0.1405	-0.0050	0.9500	1.0072	0.1338	-0.0088	0.9370	0.9617	0.1356
		0.20	0.0343	0.9221	1.3874	0.2564	-0.0005	0.9680	1.1632	0.1672	0.0026	0.9630	1.1066	0.1327	-0.0028	0.9490	1.0256	0.1309	-0.0035	0.9540	1.0070	0.1290
1000		0.01	-0.2073	0.9596	1.3802	0.5558	-0.0215	0.9590	1.0726	0.2123	-0.0100	0.9640	1.0788	0.1480	-0.0005	0.9610	1.0190	0.1394	0.0011	0.9429	0.9748	0.1389
		0.05	-0.0588	0.9219	1.4761	0.4068	-0.0200	0.9730	1.2211	0.1876	-0.0032	0.9490	1.0570	0.1510	-0.0021	0.9590	1.0346	0.1374	0.0080	0.9489	1.0037	0.1351
	1.0	0.10	0.0106	0.9230	1.4319	0.3474	-0.0341	0.9670	1.1056	0.2028	-0.0078	0.9630	1.0640	0.1492	-0.0028	0.9620	1.0507	0.1348	0.0012	0.9585	1.0235	0.1323
		0.20	0.0763	0.9262	1.3329	0.2934	-0.0133	0.9640	1.1008	0.1885	-0.0014	0.9600	1.0776	0.1442	-0.0009	0.9570	1.0227	0.1377	0.0030	0.9463	1.0062	0.1345
		0.01	-0.1018	0.9239	1.3307	0.6998	-0.1216	0.9637	0.9627	0.4050	-0.0163	0.9550	1.0197	0.2023	-0.0047	0.9580	1.0183	0.1745	-0.0013	0.9470	0.9795	0.1693
		0.05	-0.0913	0.9294	1.4332	0.5537	-0.1069	0.9558	0.9159	0.4085	-0.0206	0.9630	1.0835	0.1902	-0.0099	0.9630	1.0448	0.1700	-0.0078	0.9538	0.9944	0.1672
	2.0	0.10	0.0541	0.9113	1.3390	0.4754	-0.1020	0.9270	0.8930	0.3921	-0.0273	0.9640	1.0786	0.1911	-0.0099	0.9670	1.0377	0.1712	-0.0005	0.9599	1.0365	0.1599
		0.20	0.2376	0.8773	1.3275	0.4422	-0.0359	0.9370	0.9947	0.2988	-0.0091	0.9610	1.0572	0.1886	-0.0033	0.9690	1.0773	0.1628	0.0004	0.9567	1.0334	0.1606

Table D4: number of iterations in the simulation study that resulted in an inadmissable outcome. This mainly happened when the BCH three-step correction method was applied with low class-specific response probabilities. The inadmissable solution would here be that after correction, the posterior probabilities for all classes given a specific combination of scores would be zero, so that it was not possible to create imputations here.

			0.70			0.80			0.90			0.95			0.99		
method	Q	Z	200	500	1,000	200	500	1,000	200	500	1,000	200	500	1,000	200	500	1,000
no	0	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	0	.05	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	0	.10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	0	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	.5	.01	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	.5	.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	.5	.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	.5	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	1	.01	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	1	.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	1	.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	1	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	2	.01	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	2	.05	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	2	.10	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no	2	.20	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	0	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	0	.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	0	.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	0	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	.5	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	.5	.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	.5	.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	.5	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	1	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	1	.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	1	.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	1	.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ML	2	.01	21	2	0	11	0	0	2	0	0	6	0	0	2	0	0
ML	2	.05	13	0	0	10	1	0	4	0	0	0	0	0	1	0	0
ML	2	.10	5	0	0	3	0	0	1	0	0	0	0	0	0	0	0
ML	2	.20	2	0	0	1	0	0	5	0	0	2	0	0	1	0	0
BCH	0	.01	775	747	655	189	3	0	133	3	0	157	0	16	167	21	0
BCH	0	.05	593	371	210	46	0	0	1	0	0	12	0	0	154	60	2
BCH	0	.10	472	193	129	35	0	0	2	0	0	12	0	0	260	114	12
BCH	0	.20	384	196	110	31	0	0	5	0	0	30	0	0	462	187	9
BCH	.5	.01	779	761	691	157	7	0	138	10	0	130	4	34	169	17	1
BCH	.5	.05	634	469	290	43	2	0	6	0	0	4	0	0	134	72	0
BCH	.5	.10	450	237	212	37	1	0	4	0	0	13	0	0	225	107	0
BCH	.5	.20	421	233	178	34	1	0	5	0	0	29	0	0	467	191	0
BCH	1	.01	822	799	777	203	19	0	143	5	0	134	7	71	185	20	2
BCH	1	.05	659	597	475	53	2	0	1	0	0	10	0	0	130	74	2
BCH	1	.10	527	371	286	46	1	0	7	0	0	20	0	0	265	106	13
BCH	1	.20	473	270	255	32	0	0	3	0	0	24	0	0	406	220	13
BCH	2	.01	873	902	908	398	132	36	151	6	0	129	3	298	169	18	0
BCH	2	.05	787	759	674	212	45	4	3	1	1	12	0	1	131	73	5
BCH	2	.10	671	584	459	108	5	0	7	0	0	19	0	0	269	101	2
BCH	2	.20	521	418	340	56	0	0	2	0	0	24	0	0	435	209	8

Appendix E

The iterative method for an equality and inequality constrained minimization of a quadratic function described by Goldfarb & Idnani (1983) has been implemented in the R package `quadprog` available in the repository CRAN (Turlach & Weingessel, 2013).

The minimization procedure is implemented in the function `solve.QP` which is called as

```
solve.QP(Dmat,dvec,Amat,bvec,meq).
```

Its arguments are:

- `Dmat`: the matrix $\mathbf{D}$ appearing in the quadratic function: $(\mathbf{D}\mathbf{D}') \otimes \mathbf{I}_{n\times n}$;
- `dvec`: the vector $\mathbf{d}$ appearing in the quadratic function: $\mathbf{e}'\mathbf{P}$;
- `Amat`: The transpose of $\mathbf{H}$ ($\mathbf{H}'$) defining the linear constraints on the parameters $\mathbf{b}$;
- `bvec`: A vector of length $N+1$, with its first elements equal to 1 and the reminaing N elements all equal to 0, these are the constants $\mathbf{b}_0$ in the constraints.
- `meq`: $1+$ the number of elements in J_1

The minimization procedure can be applied using the following function:

```
qpsolve <- function(e,d,iequal){
  nr    <- nrow(e)
  nc    <- ncol(e)
  ncel  <- nr*nc
  evec  <- as.vector(e)
  id    <- diag(nr)
  p     <- kronecker(t(d),id)
  dmat  <- kronecker(d %*% t(d),id)
  dvec  <- as.vector(evec %*% p)
  im    <- diag(ncel)
  i1    <- iequal
  i2    <- setdiff(1:ncel,i1)
  index <- c(i1,i2)
  im2   <- im[index,]
  at    <- rbind(rep(1,ncel),im2)
```

```
    amat  <- t(at)
    bvec  <- c(1,rep(0,ncel))
    meq   <- 1 + length(iequal)
    res   <- solve.QP(dmat,dvec,amat,bvec,meq)
    return(res)
  }
```

The function is used by defining the **E**-matrix, the **D**-matrix and the inequality constraints:

```
res <- qpsolve(E,D,iequal).
```

In section 6.4, the extended BCH method is applied on a dataset from the Political Action Survey. There are no combinations of scores between the latent variable and the exogenous covariate that are not possible in practice, so therefore it is not needed to fix any marginals to zero. However, in this appendix a margin of the **A**-matrix is fixed for illustrative purposes.

The `qsolve()` function can be used by defining the **E**-matrix, the **D**-matrix and the inequality constraints. In the application section, the **E**-matrix, the **D**-matrix are defined, and since there are no inequality constraints, these are omitted for the function by specifying

```
iequal <- c()
```

By using the function `qpsolve(E,D,iequal)` both the unconstrained and the constrained solution for the **A**-matrix are given. The output is saved under the name `res`: `res <- qpsolve(E,D,iequal)`. The unconstrained solution can be requested by:

```
res$unconstrained.solution
```

and the constrained solution can be requested by:

```
res$solution
```

For illustration purposes, the cell $Q_{16\text{-}34} \times X_3$ of the **A**-matrix is fixed to zero. When vectorizing the **A**-matrix, this cell is the seventh element, so this needs to be specified:

```
iequal <- c(7)
```

It can now be seen that the constrained solution is not only without negative values, also the cell $Q_{16\text{-}34} \times X_3$ is fixed to zero:

$$\mathbf{A}_{\text{constraint}} = \begin{array}{c} \\ Q_{16\text{-}34} \\ Q_{35\text{-}57} \\ Q_{58\text{-}91} \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0.06007299 & 0.13800030 & 0.0000000 & 0.12215613 \\ 0.10183738 & 0.17636356 & 0.0730305 & 0.00140033 \\ 0.15732017 & 0.05457865 & 0.1152400 & 0.00000000 \end{pmatrix} \end{array}.$$

Bibliography

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* (Vol. 547). John Wiley & Sons.

André, S., & Dewilde, C. (2016). Home ownership and support for government redistribution. *Comparative European Politics*, *14*(3), 319–348. doi: 10.1057/cep.2014.31

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Using the bch method in mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, *21*(2), 1–22.

Bakk, Z. (2015). *Contributions to bias adjusted stepwise latent class modeling* (Doctoral dissertation). Retrieved from `https://pure.uvt.nl/portal/files/8521154/Bakk_Contributions_16_10_2015.pdf`

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289. doi: 10.1080/10705511.2015.1049698

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311.

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31.

Bakker, B. F., Van Rooijen, J., & Van Toor, L. (2014). The system of social statistical datasets of statistics netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, *30*(4), 411–424.

Bakker, B. F. M. (2009). Trek alle registers open! *Rede in verkorte vorm uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Methodologie van registers voor sociaalwetenschappelijk onderzoek bij de Faculteit der Sociale Wetenschappen van de Vrije Universiteit Amsterdam op 26 november 2009*. Retrieved from http://dare.ubvu.vu.nl/bitstream/handle/1871/15588/Oratie%20Bakker.pdf

Bakker, B. F. M. (2010). Micro-integration, state of the art. *Paper presented at the joint UNECE-Eurostat expert group meeting on registered based censuses in The Hague, May 11, 2010*. Retrieved from https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2010/wp.10.e.pdf

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8–17. doi: 10.1111/j.1467-9574.2011.00504.x

Barnes, S. H., Allerbeck, K. R., Farah, B. G., Heunks, F. J., Inglehart, R. F., Jennings, M. K., … Rosenmayr, L. (1979). *Political action: Mass participation in five western democracies*. Sage.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Bassi, F. (1997, Dec 01). Identification of latent class markov models with multiple indicators and correlated measurement errors. *Journal of the Italian Statistical Society*, *6*(3), 201. doi: 10.1007/BF03178912

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, *41*(1), 164–171.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys* (Vol. 568). John Wiley & Sons.

Biemer, P. P. (2004). An analysis of classification error for the revised current population survey employment questions. *Survey Methodology*, *30*(2), 127–140.

Biemer, P. P. (2011). *Latent class analysis of survey error* (Vol. 571). Hoboken, New Yersey: John Wiley & Sons.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (2011). *Measurement errors in surveys* (Vol. 173). John Wiley & Sons.

Blackwell, M., Honaker, J., & King, G. (2015). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, *46*(3), 303-341.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27.

Bos, N., Stipdonk, H., & Commandeur, J. (2017). Ernstig verkeersgewonden 2016. *SWOV Instituut voor Wetenschappelijk Onderzoek Verkeersveiligheid*. Retrieved from `https://www.swov.nl/publicatie/ernstig-verkeersgewonden-2016`

CBS. (2018a). *About us.* online. Retrieved from `https://www.cbs.nl/en-gb/about-us/organisation`

CBS. (2018b). Documentatie banen en lonen op basis van de polisadministratie. *Microdata Services*.

Census Hub. (2017, July). *European statistical system.*

Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology*, *35*(4), 1074–1081.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, *89*(428), 1314–1328.

Croon, M., Marcoulides, G., & Moustaki, I. (2002). Using predicted latent scores in general latent structure models. *Latent Variable and Latent Structure Models*, 195–224.

Daalmans, J. (2015). Estimating detailed frequency tables from registers and sample surveys. *CBS discussion paper*.

De Heer, W., & De Leeuw, E. (2002). Survey nonresponse. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), (chap. Trends in household survey nonresponse: A longitudinal and international comparison). John Wiley New York.

De Waal, T. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the IAOS*, *32*(2), 231–243. doi: 10.3233/SJI-150950

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

De Waal, T., Pannekoek, J., & Scholtus, S. (2012). The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 204–210. doi: 10.1002/wics.1194

Dewilde, C., & Decker, P. D. (2016). Changing inequalities in housing outcomes across western europe. *Housing, Theory and Society*, *33*(2), 121-161. doi: 10.1080/14036096.2015.1109545

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659. doi: 10.1007/s00180-007-0103-7

European Commission. (2008). Regulation (ec) no 763/2008 of the european parliament and of the council of 9 july 2008 on population and housing censuses. *Official Journal of the European Union*(L218), 14-20.

European Commission. (2009). Commission regulation (ec) no 1201/2009 of 30 november 2009 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns. *Official Journal of the European Union*(L329), 29-68.

European Commission. (2010). Commission regulation (eu) no 1151/2010 of 8 december 2010 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. *Official Journal of the European Union*(L324), 1-12.

Eurostat. (2012). Labour force survey in the eu, candidate and efta countries: main characteristics of the national surveys 2005.

Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association*, *71*(353), 17–35.

Filipponi, D., Guarnera, U., & Varriale, R. (2019). Hidden markov models to estimate italian employment status. *NTTS 2019 Bruxelles 11-13 March 2019*.

Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, *52*(12), 5263–5268. Retrieved from http://dx.doi.org/10.1016/j.csda.2008.04.030

Geerdinck, M., Goedhuys-van der Linden, M., Hoogbruin, E., De Rijk, A., Sluiter, N., & Verkleij, C. (2014). *Monitor kwaliteit stelsel van basisregistraties: Nulmeting van de kwaliteit van basisregistraties in samenhang, 2014* (13114th ed.). Henri Faasdreef 312, 2492 JP Den Haag: Centraal Bureau voor de Statistiek. Retrieved from https://www.cbs.nl/-/media/_pdf/2016/50/monitor-kwaliteit-stelsel-van-basisregistraties.pdf

Gerritse, S. C., Bakker, B. F., de Wolf, P.-P., & van der Heijden, P. G. (2016). Undercoverage of the population register in the netherlands, 2010. *CBS Discussion paper*, *2016*.

Goldfarb, D., & Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical programming*, *27*(1), 1–33.

Graham, A. (1982). *Kronecker products and matrix calculus with applications*. New York, NY: John Wiley & Sons Inc.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, *8*(3), 206–213.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics*, *28*(2), 173.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Guarnera, U., & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Statistical Journal of the IAOS*, *32*(4), 537-544. doi: 10.3233/SJI-150951

Hagenaars, J. A. (1993). *Loglinear models with latent variables* (Vol. 94). Sage.

Houbiers, M. (2004). Towards a social statistical database and unified estimates at statistics netherlands. *Journal of Official Statistics*, *20*(1), 55-75.

Jennings, M. K., Van Deth, J., et al. (1990). Continuities in political action. *A longitudinal study of political participation in three Western democracies*, 23–64.

Jörgren, F., Johansson, R., Damber, L., & Lindmark, G. (2010). Risk factors of rectal cancer local recurrence: population-based survey and validation of the swedish rectal cancer registry. *Colorectal Disease*, *12*(10), 977–986. doi: 10.1111/j.1463-1318.2009.01930.x

Keel, P. K., Fichter, M., Quadflieg, N., Bulik, C. M., Baxter, M. G., Thornton, L., … others (2004). Application of a latent class analysis to empirically define eatingdisorder phenotypes. *Archives of General Psychiatry*, *61*(2), 192–200.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., & Wang, Q. (2015). Simultaneous edit-imputation for continuous microdata. *Journal of the American Statistical Association*, *110*(511), 987–999. doi: 10.1080/01621459.2015.1040881

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, *40*(1), 88–104.

Kreuter, F., Yan, T., & Tourangeau, R. (2008). Good item or bad: can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(3), 723–738.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural equation modeling: a multidisciplinary journal*, *20*(1), 1–26.

Lersch, P. M., & Dewilde, C. (2015). Employment insecurity and first-time homeownership: Evidence from twenty-two european countries. *Environment and Planning A*, *47*(3), 607-624. Retrieved from dx.doi.org//10.1068/a130358p doi: 10.1068/a130358p

Linzer, D. A., Lewis, J. B., et al. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of statistical software*, *42*(10), 1–29.

Loosveldt, G. (2008). International handbook of survey methodology. In E. de Leeuw, J. Hox, & D. Dillman (Eds.), (pp. 201–220). Taylor & Francis Group New York, NY.

Magidson, J., Vermunt, J. K., & Tran, B. (2009). Using a mixture latent markov model to analyze longitudinal us employment data involving measurement error. *New trends in psychometrics*, 235–242.

Manrique-Vallier, D., & Reiter, J. P. (2013). Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, *40*, 125 - 134. Retrieved from https://ecommons.cornell.edu/handle/ 1813/34889

Manrique-Vallier, D., & Reiter, J. P. (2016). Bayesian simultaneous edit and imputation for multivariate categorical data. *Journal of the American Statistical Association*, *112*(520), 1708-1719. doi: 10.1080/01621459.2016.1231612

Manzoni, A., Vermunt, J. K., Luijkx, R., & Muffels, R. (2010). Memory bias in retrospectively collected employment carreers: a model based-approach to correct for measurement error. *Sociological methodology*, *40*(1), 39–73.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *psychometrika*, *81*(2), 274–289.

McCutcheon, A. L. (1987). Sexual morality, pro-life values, and attitudes toward abortion: A simultaneous latent structure analysis for 1978-1983. *Sociological Methods & Research*, *16*(2), 256-275. doi: 10.1177/0049124187016002003

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. doi: 10.1007/BF02294457

Mislevy, R. J., Beaton, A. E., & Kaplan, a., B. (1992, June). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161. doi: 10.1111/j.1745-3984.1992.tb00371.x

Monseur, C., & Adams, R. (2009, January). Plausible values: How to deal with their limitations. *Journal Of Applied Measurement*, *10*(3), 320-334. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19671992

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Officiel Statistics*, *16*(4), 331–362.

Mulder, C. H. (2006). Home-ownership and family formation. *Journal of Housing and the Built Environment*, *21*(3), 281–298. doi: 10.1007/s10901-006-9050-9

National Office for Identity Data, Ministry of the Interior and Kingdom Relations. (2016, November). *The dutch brp register: registration of personal details for the government and for you.* Retrieved from https://www.government.nl/binaries/government/documents/leaflets/2017/01/19/brochure-brp-engelstalig/Brochure+BRP+-+Engelstalig+-+def+versie+voor+publicatie+lowres.pdf

Ness, A. R. (2004). The Avon Longitudinal Study of Parents and Children (ALSPAC)–a resource for the study of the environmental determinants of childhood obesity. *European journal of endocrinology*, *151*(Suppl 3), U141–U149. doi: 10.1530/eje.0.151U141

Nordholt, E. S., Hartgers, M., & Gircour, R. (2004). The dutch virtual census of 2001. *The Hague/Heerlen: Statistics Netherlands*.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782–797.

Oberski, D. L. (2015). Total survey error in practice. In P. P. Biemer et al. (Eds.), *Total survey error* (chap. 16 Estimating error rates in an administrative register and survey questions using a latent class model). New York: Wiley.

Pankowska, P., Bakker, B., Oberski, D., & Pavlopoulos, D. (n.d.). How linkage error affects hidden markov models.

Pankowska, P., Bakker, B., Oberski, D., & Pavlopoulos, D. (2017a, 3). Estimating employment mobility using linked data from different sources. does linkage error matter? In *Ntts 2017.*

Pankowska, P., Bakker, B., Oberski, D. L., & Pavlopoulos, D. (2017b). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. *Statistical Journal of the IAOS*(Preprint), 1–13.

Pankowska, P., Pavlopoulos, D., Oberski, D., & Bakker, B. (2018). Dependent interviewing: a remedy or a curse for measurement error in surveys? In *International total survey error workshop.* Retrieved from `https://dism.ssri.duke.edu/sites/dism.ssri.duke.edu/files/pdfs/itsew_program_final.pdf`

Pavlopoulos, D., & Vermunt, J. (2015). Measuring temporary employment. do survey or register tell the truth? *Survey Methodology*, *41*(1), 197-214.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*(480), 1462–1471.

Reurings, M. C. B., & Bos, N. M. (2012). Ernstig verkeersgewonden in de jaren 2009 en 2010: update van de cijfers. *SWOV Stichting Wetenschappelijk Onderzoek Verkeersveiligheid*.

Reurings, M. C. B., & Stipdonk, H. L. (2009, December). Ernstig gewonde verkeersslachtoffers in nederland in 1993-2008. *Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV*. Retrieved from https://www.swov.nl/publicatie/ernstig-gewonde-verkeersslachtoffers-nederland-1993-2008

Reurings, M. C. B., & Stipdonk, H. L. (2011). Estimating the number of serious road injuries in the netherlands. *Annals of epidemiology*, *21*(9), 648–653.

Robertsson, O., Dunbar, M., Knutson, K., Lewold, S., & Lidgren, L. (1999). Validation of the swedish knee arthroplasty register: A postal survey regarding 30,376 knees operated on between 1975 and 1995. *Acta Orthopaedica Scandinavica*, *70*(5), 467-472. doi: 10.3109/17453679909000982

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: how the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *109*(1), 56–61. doi: 10.1177/0759106310387713

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2014). Predictive inference using latent variables with covariates. *Psychometrika*, 283–314.

Scholtus, S. (2009). Automatic detection of simple typing errors in numerical data with balance edits. *CBS discussion paper*. Retrieved from https://www.cbs.nl/-/media/imported/documents/2009/48/2009-46-x10-pub.pdf

Scholtus, S. (2011). Algorithms for correcting sign errors and rounding errors in business survey data. *Journal of Official Statistics*, *27*(3), 467.

Scholtus, S., & Bakker, B. F. (2013). Estimating the validity of administrative and survey variables through structural equation modeling. *CBS discussion paper*.

Schott, J. R. (1997). *Matrix analysis for statistics*. John Wiley & Sons.

Schrijvers, C. T. M., Stronks, K., van de Mheen, D. H., Coebergh, J.-W. W., & Mackenbach, J. P. (1994). Validation of cancer prevalence data from a postal survey by comparison with cancer registry records. *American Journal of Epidemiology*, *139*(4), 408-414. doi: 10.1093/oxfordjournals.aje.a117013

Schulte Nordholt, E., Van Zeijl, J., & Hoeksma, L. (2014). Dutch census 2011, analysis and methodology. *Statistics Netherlands*. Retrieved from `https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf`

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499–521. doi: 10.3102/1076998613480394

Singh, A. C., & Rao, J. (1995). On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, *90*(430), 478–488.

Spiegelman, D., McDermott, A., & Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American journal of clinical nutrition*, *65*(4), 1179S–1186S.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Tempelman, C. (2007). *Imputation of restricted data: applications to business surveys* (Doctoral dissertation, Rijksuniversiteit Groningen). Retrieved from `https://www.cbs.nl/-/media/imported/documents/2007/05/2007-i76-pub.pdf`

The Economic and Social Council. (2005). Ecosoc resolution 2005/13. *2010 World Population and Housing Census Programme*. Retrieved from `http://www.un.org/en/ecosoc/docs/2005/resolution%202005-13.pdf`

Turlach, B. A., & Weingessel, A. (2013). *quadprog: Functions to solve quadratic programming problems. r package version 1.5-5.*

Turner, C. F., Smith, T. K., Fitterman, L. K., Reilly, T., Pate, K., Witt, M. B., ... Forsyth, B. H. (1997). The quality of health data obtained in a new survey of elderly americans: A validation study of the proposed medicare beneficiary health status registry (mbhsr). *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52B*(1), S49-S58. doi: 10.1093/geronb/52B.1.S49

Understanding Society. (2016). Understanding society: Innovation panel, waves 1-7, 2008-2014. [data collection]. 6th edition [Computer software manual]. UK Data Service. doi: 10.5255/UKDA-SN-6849-7

University of London. UCL Institute of Education. Centre for Longitudinal Studies, Millennium Cohort Study: First Survey, 2001-2003 [computer file]. 6th Edition. Colchester, Essex: UK Data Archive [distributor], SN: 4683. (2007, March).

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, 1–21. doi: 10.1007/s00357-016-9195-5

Van der Vaart, W., & Glasner, T. (2007). Applying a timeline as a recall aid in a telephone survey: a record check study. *Applied Cognitive Psychology*, *21*(2), 227–238. doi: 10.1002/acp.1338

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2000). Graphical displays for latent class cluster and latent class factor models. In *Proceedings in computational statistics* (pp. 121–122).

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 549–553. Retrieved from http://members.home.nl/jeroenvermunt/ermss2004a.pdf

Vermunt, J. K., & Magidson, J. . (2013a). Latent GOLD 5.0 Upgrade Manual [Computer software manual]. Belmont, MA. Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LG5manual.pdf

Vermunt, J. K., & Magidson, J. (2013b). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. *Statistical Innovations Inc., Belmont, MA*. Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf

Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*(1), 369–397. doi: 10.1111/j.1467-9531.2008.00202.x

Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv preprint arXiv:1409.8542*. Retrieved from https://arxiv.org/abs/1409.8542

Wang, C.-P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models. *Journal of the American Statistical Association*, *100*(471), 1054-1076. doi: 10.1198/016214505000000501

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & Group, N. P. S. D. W. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. doi: 10.1111/j.1472-4642.2008.00482.x

Wong, E. (2011). Abbreviated injury scale. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 5–6). New York, NY: Springer New York.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114-128. doi: 10.1016/j.stueduc.2005.05.005

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. doi: 10.1111/j.1467-9574.2011.00508.x

Zhang, L.-C., & Pannekoek, J. (2015). Optimal adjustments for inconsistency in imputed data. *Survey Methodology*, *41*(1), 127–144. Retrieved from http://www.statcan.gc.ca/pub/12-001-x/12-001-x2015001-eng.pdf

Samenvatting in het Nederlands

Bij het produceren van officiële statistieken, maakt het CBS zoveel mogelijk gebruik van bestaande administratieve registers. Echter, niet alle variabelen die nodig zijn voor de productie van officiële statistieken zijn beschikbaar in deze registers, en daarom worden deze registers aangevuld met informatie verkregen door middel van surveys. Wanneer het CBS officiële statistieken produceert, zijn er een aantal zaken waar rekening mee moet worden gehouden. Ten eerste mogen de statistieken geen combinaties van scores bevatten die in de praktijk onmogelijk zijn. Zo is bijvoorbeeld de combinatie van 'leeftijd=jonger dan 5 jaar' en 'burgerlijke staat = gehuwd' niet mogelijk omdat dit wettelijk verboden is. Ten tweede moeten alle geproduceerde statistieken consistent zijn. Dit betekent dat wanneer een kruistabel tussen de variabelen 'opleidingsniveau $\times$ geslacht $\times$ regio' wordt geproduceerd, en daarnaast ook een kruistabel 'opleidingsniveau $\times$ geslacht $\times$ burgerlijke staat', dat, bijvoorbeeld, het totaal aantal hoogopgeleide mannen in beide kruistabellen exact gelijk aan elkaar moet zijn. Dit kan een probleem zijn wanneer een variabele (zoals bijvoorbeeld 'geslacht') onafhankelijk gemeten is in verschillende registers en surveys en meetfout bevat, waardoor waarden gemeten in deze registers en surveys van elkaar kunnen afwijken.

Om gelijktijdig met deze twee problemen om te gaan wordt in dit proefschrift voorgesteld om gebruik te maken van een combinatie van latente klassenanalyse en multipele imputatie. Bij een latente klassenmodel kunnen verschillende variabelen die hetzelfde meten gebruikt worden als indicatoren van een latente variabele die de 'werkelijke scores' van deze variabele meet. Wanneer er dan sprake is van meetfout in één van de indicatoren kan door middel van dit model een inschatting worden gemaakt van de juiste score voor deze persoon. Ten tweede kunnen onmogelijke combinaties van scores expliciet gespecificeerd worden in het latente klassenmodel en kunnen zo worden voorkomen. Door vervolgens gebruik te maken van multipele imputatie kunnen consistente statistieken worden geproduceerd en kan onzekerheid door ontbrekende en conflicterende waarden correct worden weergegeven.

In dit proefschrift wordt ten eerste bovenstaande methode geïntroduceerd. Daarnaast wordt de methode uitgebreid voor verschillende specifieke situaties, zoals meerdere latente variabelen en longitudinale data. De methode wordt ook zo uitgebreid dat covariaten op een later tijdstip aan het model toegevoegd kunnen worden. Ten slotte wordt onderzocht of en hoe de methode gebruikt kan worden voor de productie van populatiestatistieken.

Dankwoord

Sommigen van jullie denken misschien dat dit proefschrift tot stand is gekomen terwijl ik in m'n eentje vier jaar lang driftig op mijn laptop heb zitten tikken, maar niets is minder waar. Een hoop mensen hebben hier een grote rol in gehad en ik wil hen hier ontzettend voor bedanken.

Om te beginnen met mijn begeleiders. Ton, ik heb niet alleen veel van je geleerd over editen en imputeren, maar ook over hoe de wereld van de officiële statistiek in elkaar zit. Ik sprak je vaker op de gang, bij de koffieautomaat of op de flexplekken van het CBS dan tijdens ingeplande afspraken, waardoor ik altijd het gevoel had heel zelfstandig bezig te zijn terwijl je toch constant een oogje in het zeil hield. Bovenal hebben we samen veel grappen gemaakt, geklaagd en gedaan alsof we dom en lui zijn, maar we weten allebei wel beter.

Daniel, vooral in het begin heb je me veel begeleiding gegeven. Ik was toen behoorlijk overweldigd door je gigantische hoeveelheid kennis en je constante stroom van ideeën. Je was ook wel streng. Zo liet je me de introductie van het eerste paper zeven keer opnieuw schrijven, maar hierdoor kon ik latere projecten een stuk zelfstandiger aanpakken. Doordat je in Utrecht ging werken werd onze samenwerking wat minder intensief, dus ik ben blij dat we de komende jaren weer samen aan de slag gaan met je nieuwe plannen!

Jeroen, ik was al een tijdje met latenteklassenmodellen bezig toen je je intensiever met mijn project ging bezighouden. Ik heb in deze periode nog veel van je mogen leren over dit onderwerp: van Latent GOLD trucjes tot oplossingen voor allerlei problemen die je 'eigenlijk ook gewoon als een mixture kunt zien'. Juist doordat je er iets verder vanaf stond heb ik door onze gesprekken geleerd zelf ook kritisch over de relevantie van verschillende onderdelen van het project na te blijven denken.

Ik wil de commissie bedanken voor het lezen en beoordelen van dit proefschrift en dat zij de reis naar Tilburg hebben willen maken om bij deze verdediging aanwezig te zijn. Barry en Frank, bedankt dat ik jullie gecombineerde dataset mocht gebruiken. hierdoor kon ik het tweede hoofdstuk vlot afronden. SWOV, bedankt voor de toegang tot jullie data over ernstige verkeersgewonden die ik heb gebruikt voor hoofdstuk drie. Roberta and Danila, thank you for your contributions to the fourth chapter of this dissertation, but even more for being wonderful hosts during my stay at ISTAT. When I look back, it's unbelievable how much we've accomplished in just three months and while having so much fun. Marcel, onze samenwerking begon met een praktisch probleem waar ik tijdens hoofdstuk vijf tegenaan liep, maar door je enthousiasme en uitgebreide kennis over dit onderwerp hebben we er

uiteindelijk een heel extra hoofdstuk over geschreven. Bedankt! Sander en Jacco, op het CBS bespraken we onze gemeenschappelijke onderzoeksinteresses al vaak, en ik ben blij dat hier een samenwerking uit voort is gekomen in de vorm van hoofdstuk zeven.

Ik heb het altijd heel erg naar mijn zin gehad in Tilburg, en hiervoor wil ik ten eerste mijn kamergenoten bedanken. Andrea, Eva, Davide en Sara, onze kamer was altijd een gezellige plek om te kletsen en klagen, maar ook om elkaar te helpen met praktische of onderzoeksgerelateerde zaken. Daarnaast heb ik de (vele) treinuren, lunches, gesprekken en borrels ook altijd erg gezellig gevonden met collega's. Bedankt Chris, Elise, Erwin, Esther, Felix, Ijsbrand, Inga, Jesper, Jules, Leonie, Marcel, Mattis, Michèle, Niek, Paul, Paulette, Robbie, Tim, Soogeon en alle andere afdelingsgenoten. Ik wil de Vici en later de Psychometrics groep bedanken voor jullie nuttige commentaar en de mogelijkheid om ook over jullie werk te leren. Anne-Marie en Marieke, bedankt dat ik altijd voor van alles bij jullie binnen kon lopen! Arnout, Bart en Nino, bedankt voor de prettige samenwerking en de gezelligheid op het CBS.

Gedurende mijn periode in Tilburg ben ik steeds betrokken geweest bij onderwijs. Eerst MTO-E, dit was op veel verschillende manieren een uitdaging, maar altijd gezellig. Bedankt daarvoor John, Leonie, Hilde, Pia en Mattis. Jelte, bedankt dat je me de kans hebt gegeven om MTO-B te coördineren en een extra research visit te maken. Guy, Ghislaine, Noémi en Wilco bedankt dat ik altijd bij jullie binnen kon lopen met vragen over MTO-B. Joris, bedankt dat je zo'n chille buddy was!

Ik wil graag de studenten en docenten van de M&S-afdeling in Utrecht bedanken voor de fijne tijd tijdens de research master. Bente, Danielle, Eva, Jolien, Kees, Kirsten, Mariëlle, Rob, Sanne en Timo ik ben blij dat we elkaar nog altijd spreken. Joop, Edith en Gerko, jullie hebben me zo goed opgevangen tijdens het tweede jaar en de samenwerking was altijd zo prettig, gezellig en leerzaam. Die gezelligheid vond ik ook terug bij de vele survey-, officiële statistiek- en IOPS-conferenties. Hiervoor wil ik in het bijzonder Anne, Paulina en Peter bedanken.

Eva en Wies, wat fijn dat jullie vandaag achter me staan. Eva, vanaf het moment dat je hoorde dat ik ook in Tilburg kwam werken, heb je altijd voor me klaargestaan. Zo regelde je al dat ik niet 'in Enschede' kwam te zitten voordat ik überhaupt wist wat dat betekende. Wies, twaalf jaar geleden vielen we allebei flink uit te toon op een feest dat geen plaats verdient in dit boek, maar sindsdien zijn we beste vriendinnen. Ik kon de afgelopen jaren soms erg druk zijn met werk, maar jij zorgde ervoor dat er altijd leuke dingen in mijn agenda bleven staan.

Anne, Jeanine, Katja, Karin, Kim, Lucie, Melissa en Wies, ondanks dat we allemaal wat ouder worden en onze levens ook wat serieuzer, lijkt het alsof er sinds de middelbare school niks is veranderd als we bij elkaar zijn en ik hoop dat dat nog lang zo zal blijven. Zolang we onszelf de ***jes blijven noemen zit dat denk ik wel goed. Karin, bedankt dat je wat kleur aan dit proefschrift hebt gegeven. Eva en Roline, ik ben blij drie van mijn favoriete dingen met jullie te kunnen delen: statistiek, Den Haag en saté! Mai en Yvette, wat fijn dat we elkaar na zoveel jaren nog steeds zien, 'helaas' iets vaker om te lunchen of eten dan bij bootcamp. Lentine, ik denk met veel plezier en wat plaatsvervangende schaamte terug aan onze eerste onderzoeksprojecten bij politicologie en ben blij dat we die herinneringen nog regelmatig ophalen. Tess, als ik weer eens op reis ben naar een onbesuisde bestemming, kan het zomaar gebeuren dat jij er ook bent om me de leukste plekjes te laten zien en dan voelt de wereld weer heel klein.

Mama, toen ik in Tilburg begon kwam ik nog elke week bij je logeren. Inmiddels zou je daar helemaal geen zin en tijd meer voor hebben met je drukke sociale leven, en daar ben ik eigenlijk heel blij mee! Kars, het is supertof om te zien hoe jij je eigen pad volgt en ik ben jaloers op alle mooie reizen die je daarbij maakt. Oma en Volkert, de Vuntuslaan zal altijd de meest ontspannen plek ter wereld blijven.

Tiem, de jaren waarin dit proefschrift is geschreven hebben ook jouw leven erg beïnvloed. Je moest rekening houden met mijn belachelijke reis- werkschema, je hielp me allerlei R problemen op te lossen, formuleerde nette Engelse zinnen voor me en zette honderden cappucino's. Bedankt hiervoor, maar eigenlijk nog veel meer voor alle niet-werkgerelateerde dingen.

Papa, wat vind ik het jammer dat je er niet bij bent vandaag. Het bier en de bitterballen hadden je vast goed gesmaakt. Ik mis je en ik hoop dat je trots bent.